EFFICIENT ALIGNMENT-FREE SOFTWARE APPLICATIONS FOR NEXT GENERATION SEQUENCING-BASED MOLECULAR EPIDEMIOLOGY

A Dissertation
Presented to
The Academic Faculty

by

Hector Fabio Espitia Navarro

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
May 2020

EFFICIENT ALIGNMENT-FREE SOFTWARE APPLICATIONS FOR NEXT GENERATION SEQUENCING-BASED MOLECULAR EPIDEMIOLOGY

Approved by:

Dr. I. King Jordan, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Srinivas Aluru
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Jung Choi
School of Biological Sciences
Georgia Institute of Technology

Dr. Leonard Mayer
School of Medicine
Emory University

Dr. Lavanya Rishishwar
School of Biological Sciences
Georgia Institute of Technology

Date Approved: December 9, 2019

To my beloved family and my father's memory

ACKNOWLEDGEMENTS

I must thank sincerely to my advisor Dr. King Jordan for his guidance and support. I always will be grateful to him for giving me the opportunity to work in his lab. His great ability to synthesize complex ideas and communicating science have always impressed me and have been an inspiration during my time working with him.

I am also very thankful to Dr. Lavanya Rishishwar for his guidance and mentoring during my time as a graduate student at the Jordan Lab. His expertise, ideas, and suggestions have helped me greatly in my research.

I am grateful to Dr. Leonard Mayer, Dr. Srinivas Aluru, and Dr. Jung Choi. I am very fortunate to have had them in my research committee and I truly appreciate the guidance and insightful ideas they provided for my research.

I want to thank my colleagues and friends from the Jodan Lab: Aroon Chande for his valuable collaboration, and Luz Medina and Shashwat Deepali Nagar for their close friendship and good moments during my time at the laboratory.

I feel truly grateful to Yisel Carrillo for her love, patience, and support during this journey of my PhD. She always has been here to me even when we were far away from each other.

Lastly, I want to express my most profound gratitude to my mother Alba Navarro and my brother Rubén Darío Espitia for their unconditional love and support. I would not have been able to pursue my PhD without their sacrifice and endless encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AMR	Antimicrobial resistance
ARG-ANNOT	Antibiotic resistance gene-annotation
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
CARD	Comprehensive antibiotic resistance database
CDC	Centers for Disease Control and Prevention
CGE	Center for Genomic Epidemiology
cgMLST	Core genome MLST
eMLST	Extended MLST
ESA	Enhanced suffix array
MLST	Multilocus sequence typing
NGS	Next generation sequencing
PCR	Polymerase chain reaction
PFGE	Pulsed-field gel electrophoresis
RAM	Random access memory
RFLP	Restriction fragment length polymorphisms
rMLST	Ribosomal MLST
SNV	Single nucleotide variant
SRA	Sequence Read Archive
ST	Sequence type
STEC	Shiga toxin producing *Escherichia coli*

VF	Virulence factor
VFDB	Virulence factor database
wgMLST	Whole genome MLST
WGS	Whole genome sequencing

SUMMARY

Public health agencies increasingly couple next generation sequencing (NGS) based characterization of microbial genomes with bioinformatics analysis methods for molecular epidemiology. The overhead associated with the bioinformatics methods used for this purpose, in terms of both the required human expertise and computational resources, represents a critical bottleneck that limits the potential impact of microbial genomics on public health. This is particularly true for local public health agency laboratories, which are typically staffed with microbiologists who may not have substantial bioinformatics expertise or ready access to high-performance computational resources. There is a pressing need for bioinformatics solutions to genome-enabled molecular epidemiology that is accurate, easy to use, fast, and computationally efficient.

The development of an alignment-free algorithm for NGS data analysis and its implementation into turn-key software applications tailored explicitly for genome-enabled molecular epidemiology and environmental microbial genomics is the focus of my research. I explored a computational strategy based on *k*-mer frequencies to distinguish among sequences of interest in NGS read samples. By combining this strategy with the efficient data structure enhanced suffix array (ESA), I developed a base algorithm for the rapid analysis of unprocessed NGS reads. I further adapted and implemented this algorithm into a suite of software applications for sequence typing, gene detection, and gene-based taxonomic read classification.

My thesis research focused on three specific aims: (1) development of an alignment- and assembly-free algorithm and software solution for NGS-based molecular

epidemiology, (2) development of an alignment- and assembly-free fully automated Web-platform for the comprehensive characterization of bacterial isolates using whole genome sequencing (WGS) data; and (3) expanding the applicability of the alignment-free algorithm to different problems.

Sequence typing and gene detection are essential for pathogen characterization in genome-enabled approaches to molecular epidemiology. In this sense, I developed an assembly- and alignment-free algorithm, STing, which I implemented into two turn-key software utilities for sequence typing, and gene detection. Benchmarking and validation analyses showed that STing is an ultrafast and accurate solution for genome-enabled molecular epidemiology, which performs better than existing bioinformatics methods for sequence typing and gene detection.

Limited access to bioinformatics-related infrastructure and expertise impedes the successful adoption of genome-enabled approaches to molecular epidemiology in public health. To overcome this challenge, I developed WebSTing, a Web-platform that uses the STing algorithm to supply easy access to the accurate and rapid alignment-free automated characterization of WGS samples of bacterial isolates.

To demonstrate the utility of STing in problems beyond simple sequence typing and gene detection, I applied the alignment-free algorithm to two different areas: (1) public health, with the virulence gene profiling of Shiga toxin-producing *Escherichia coli* (STEC) isolates, and (2) environmental microbial genomics, with the *nifH* gene-based taxonomic classification of amplicon sequencing reads. I showed that STing performs better than the gold-standard method for STEC isolate characterization and that it correctly classifies amplicon sequencing reads on simulated communities of nitrogen-fixing organisms.

Research advance 1*:* A novel *k*-mer frequencies approach combined with the ESA data structure was used to develop an assembly- and alignment-free algorithm, STing, for rapid analysis of unprocessed NGS reads. The STing algorithm was implemented into two software applications for genome-enabled molecular epidemiology: (1) the STing typer for sequence typing, and (2) the STing detector for gene detection. The STing typer utility was compared to six widely used programs for genome-enabled sequence typing, using the traditional multilocus sequence typing (MLST) scheme, and two larger typing schemes, ribosomal MLST (rMLST), and core genome MLST (cgMLST). Comparison results showed that STing outperformed the other applications in terms of accuracy and efficiency (runtime and RAM) using the MLST and rMLST schemes and was second with the cgMLST scheme. Most importantly, STing was the only application able to perform the typing analysis using all three of the typing schemes assessed, while also showing the ability to scale successfully to genome-enabled typing schemes like cgMLST. The detector utility was used to evaluate the ability of STing for detecting two epidemiologically relevant types of markers: antimicrobial resistance (AMR) genes and virulence factor (VF) genes. Results showed that STing had 100% accuracy in detecting AMR and VF genes on 71 WGS samples generated from 17 bacterial species of high priority in clinical microbiology research.

Research advance 2: A Web-based platform, WebSTing, was developed to provide fully automated NGS-based characterization of bacterial pathogens. The main goal of WebSTing to provide easy access to genome-enabled approaches to molecular epidemiology in public health laboratories that have limitations in bioinformatics-related infrastructure and expertise. WebSTing uses the STing algorithm and supplies assembly- and alignment-free sequence typing, gene detection, and phylogenetic analysis of WGS samples of bacterial isolates.

Research advance 3: The applicability of the STing algorithm was expanded to solve problems in two different areas: (1) public health, and (2) environmental microbial genomics. For the public health application, STing was used for virulence gene profiling STEC isolates from WGS samples. Here, STing was compared to the PCR method, the current gold-standard technique used by public health laboratories for STEC characterization. Results showed that STing is more accurate than PCR for characterizing virulence genes in STEC samples. STing showed between 98% and 100% accuracy in characterizing the four genes used as markers for STEC determination (*stx1*, *stx2*, *eae*, and *ehxA*), compared to a PCR accuracy between 90% and 94%. Most importantly, and unlike the PCR technique, STing was able to detect novel genes in the analyzed STEC isolates. In the environmental microbial application, the STing algorithm was extended for *nifH* gene-based taxonomic classification of amplicon sequencing reads. The algorithm was implemented into the STing classifier utility. Using a *nifH* gene reference database, the STing classifier program was able to classify reads correctly in samples of nitrogen-fixing organism communities, simulated up to the lowest sequencing depth of 1x coverage. Importantly, results showed that full-length reference sequences of the gene *nifH* led to better taxonomic classification of the sequencing reads.

CHAPTER 1. INTRODUCTION – RAPID BACTERIAL TYPING IN THE POST-GENOMIC ERA: DEVELOPMENTS IN COMPUTATIONAL METHODS

1.1 Molecular epidemiology and typing

Epidemiology entails the study of population distributions of determinants of health and disease, and molecular approaches to epidemiology rely on the analysis of genetically encoded biomarkers and risk factors (Wang, et al., 2015). Molecular epidemiology studies are critically important for public health surveillance as well as disease management and control. In the post-genomic era, which is characterized by the rapid accumulation of numerous whole genome sequences, molecular epidemiology increasingly relies on genome-enabled techniques. Genomic approaches to molecular epidemiology necessitate the use of sophisticated computer algorithms capable of analyzing massive amounts of data for the presence and distribution of genetic markers and risk factors. In this chapter, we cover the state-of-the-art with respect to the computational genomic approaches used to support molecular epidemiology and typing.

Molecular typing refers to the identification of the specific 'types' of microbial pathogens that cause infectious disease. For the most part, this concerns the set of procedures used to identify distinct strains of bacteria within a given species. Accordingly, molecular typing techniques require a high level of resolution in order to distinguish very closely related organisms, which is critically important for molecular epidemiology (Wang, et al., 2015). The accurate identification and discrimination of bacterial strains within a given pathogenic species allows scientists to: (i) address the underlying biology of bacterial pathogenicity, including virulence, transmissibility, and response to drugs and vaccines, (ii) track the spread of bacterial pathogens locally and globally, (iii) identify natural hosts for bacterial pathogens and associate them with specific outbreaks, and (iv) infer the evolution and population structure of bacterial pathogens. The fundamental knowledge gained from the molecular typing of bacterial pathogens facilitates the design of public health strategies for the control and prevention of infectious disease, including tailored treatment schemes, vaccine development, and vaccine surveillance programs.

Early approaches to molecular typing employed a wide variety of surrogate techniques that allowed for the indirect study of genetic variation among bacterial pathogens. These surrogate techniques measured the properties of bacterial proteins or cell surface antigens, via Western or immunoblotting and serotyping for example, or nucleic acids assayed via non-sequencing based techniques, such as restriction fragment length polymorphisms (RFLP) or polymerase chain reaction (PCR). While the development and application of these early molecular techniques provided an important advance in bacterial typing, they were difficult to standardize, replicate, and scale-up. Perhaps most importantly, surrogate techniques for molecular typing did not yield the depth of resolution needed to unambiguously distinguish closely related strains within multiple species of bacterial pathogens. The introduction of genetic sequence-based techniques for molecular typing provided a quantum leap in terms of resolution, stability, and reproducibility for the typing of bacterial pathogens.

1.1.1 *Multilocus sequence typing (MLST)*

The first *bona fide* gene sequence based technique developed for bacterial typing is referred to as multilocus sequence typing (MLST). MLST was developed by the group of Martin Maiden at Oxford University for the analysis of *Neisseria meningitidis* and was intended to be a so-called ‘portable’ typing scheme with results that could be directly compared among different laboratories around the world (Maiden, et al., 1998). It should be noted that the sequencing and analysis of 16S ribosomal RNA genes (or 16S rRNA) has also been widely used for the characterization of the evolutionary relationships among bacterial species and pre-dates MLST by more than 20 years. However, 16S rRNA sequencing typically does not provide sufficient resolution for the discrimination of distinct strains within bacterial species. Indeed, Maiden and colleagues have provided an overview of the resolution of a variety of sequence-based typing schemes and show that 16S rRNA sequence analysis provides the most reliable resolution at the level of bacterial genus and above (Maiden, et al., 2013).

MLST employs typing schemes that are specifically tailored for individual bacterial species. Species-specific MLST typing schemes rely on sequencing fragments of a set of housekeeping genes, typically 7-9 loci, which are distributed around the genome. Essential housekeeping genes are chosen for MLST to ensure that the loci are universally present among isolates that are to be typed. Distinct gene sequences for each locus in an MLST scheme are referred to as alleles, and differences between alleles across all loci in the scheme are used to distinguish specific types (or strains) of bacteria within a species. Each distinct sequence (allele) of a given MLST locus is identified by a gene (locus) name and an integer number that uniquely identifies the allele. Locus-specific integer numbers denote the order of discovery for the alleles at that locus. For example, the ABC transporter ATP-binding gene *abcZ* is one of 7 loci used as part of the traditional *N. meningitidis* MLST scheme; unique alleles of *abcZ* are denoted as *abcZ_1*, *abcZ_2*, etc., and as of this writing 881 distinct *abcZ* alleles have been identified in *N. meningitidis*. The combination of alleles characterized across all loci of the MLST scheme defines an allelic profile which is labeled with an arbitrary number that identifies a sequence type (ST). For example, for *N. meningitidis*, the combination of the alleles *abcZ_1*, *adk_3*, *aroE_4*, *fumC_7*, *gdh_1*, *pdhC_1*, and *pgm_3*, results in the allelic profile 1-3-4-7-1-1-3 that represents sequence type 2 (ST2) (Figure 1). Each species-specific MLST scheme uses a database that contains all the known alleles for each locus in the scheme and a table that associates each observed allelic profile with a ST. To characterize an isolate, the seven loci of the scheme of the species under study are sequenced, and each locus-specific sequence is compared to the allele database of the scheme, using a sequence similarity search program such as BLAST+ (Camacho, et al., 2009), to generate the allelic profile of the isolate. Finally, the unique ST identifier for the isolate is retrieved from the table of allelic profiles. STs for multiple isolates can be compared, using a minimum spanning tree for example (Figure 1), to get a sense of the scope of diversity found in a given study.

MLST was introduced in 1998, about six years prior to start of the next-generation sequence (NGS) revolution. At that time, sequencing was done using the Sanger method, which despite numerous technological improvements over the years was still relatively low-throughput, labor-intensive, time-consuming, and expensive. Given the technological limitations at the time, MLST was designed in such a way as to capture genome-wide patterns of sequence variation via sequencing a very small portion of the entire genome.

For instance, MLST alleles in the original *N. meningitidis* typing scheme are approximately 450bp long per locus. The total length of the seven allele sequences in this scheme is 3,284bp, which represents a mere ~0.1% of an entire 2.3Mbp *N. meningitidis* genome sequence. It is quite remarkable to consider how successful MLST has been for (fairly) high resolution bacterial typing given the diminishingly small percentage of overall genome sequence diversity that is represented in each scheme.

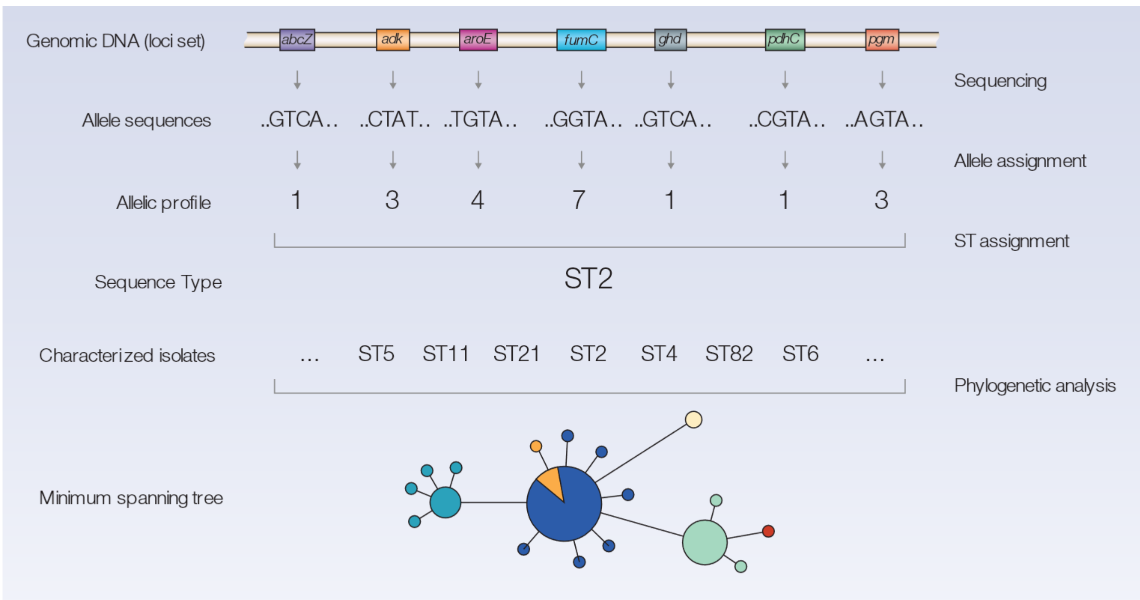


Figure 1. Graphic representation of the multilocus sequence typing (MLST) method. An example is shown for the traditional MLST scheme used for Neisseria meningitidis. Seven different loci, distributed around the genome (not shown to scale), are used for this scheme. Unique allele sequences for each locus are characterized and compared against a species-specific MLST database to yield an allelic profile, and each allelic profile is then associated with a specific sequence type (ST). Multiple STs from one or more studies can be compared using phylogenetic analyses to characterize the extent of diversity and relationships seen among a set of bacterial isolates.

One way that MLST was scaled-up was through the use of 96-well plates to perform multiple simultaneous PCRs for specific amplicons across different bacterial isolates. PCR products were then characterized using Sanger sequencing reactions and analyzed on a parallel capillary electrophoresis instrument. MLST software packages, first STARS and later MGIP, were then used to automatically convert Sanger sequencing chromatograms to allele calls and sequence types (Katz, et al., 2009). Further extensions of MLST were developed by including additional loci, particularly more variable antigen encoding loci, to yield so-called MLST+ or extended MLST (eMLST) schemes. Extended schemes for *N. meningitidis* typically include combinations of an additional six loci, including the *porA*, *porB*, *fHbp* and *fetA* antigen encoding genes. The inclusion of antigen encoding genes not only provides additional resolution to traditional MLST schemes, but can also yield valuable information with respect to vaccine design and measurement of response.

1.2 Impact of NGS on bacterial typing schemes

The advent of NGS techniques, and the resulting explosion of bacterial genome sequences (Figure 2), has led to the development of new genome-enabled approaches for bacterial typing. First and foremost, it quickly became faster and more cost effective to sequence an entire genome of a bacterial isolate using NGS platforms (initially Roche 454 and now primarily Illumina) than to amplify multiple specific MLST loci and perform Sanger sequencing on individual amplicons. Whole genome sequencing obviously yields a massive amount of data far in excess of what is provided by traditional 7-9 loci MLST schemes. This explosion of sequence data presented two distinct opportunities for bacterial typing, each of which came with its own set of computational challenges: (1) the use of whole genome sequence data for existing MLST schemes, and (2) the development of novel, larger-scale typing schemes, which avail themselves of the substantial data generated by NGS. We will cover these two broad technological developments in turn, with an emphasis on the computational approaches used for each.

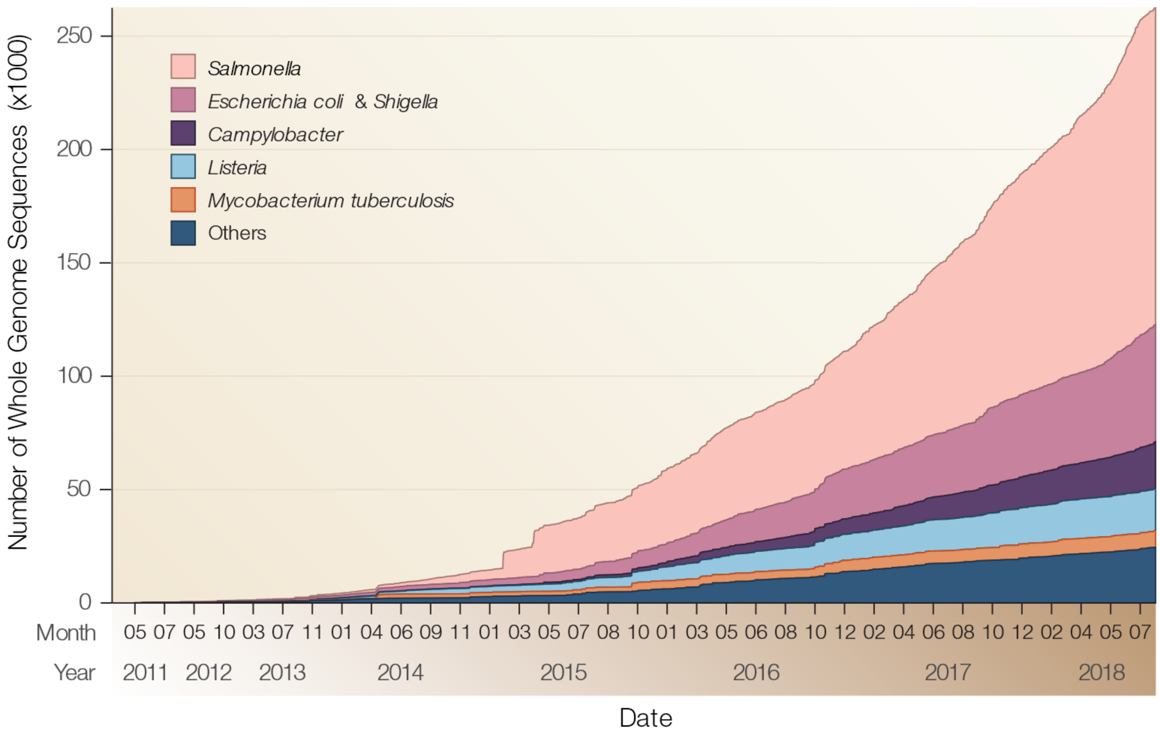


Figure 2. Growth in whole genome sequencing (WGS) of bacterial pathogens in the last seven years. The graph represents the number of WGS data submitted to NCBI's Pathogen Detection database since 2011.

Given the ability to readily generate whole genome sequences via NGS, one may wonder why a small-scale approach like MLST would be needed at all. It may seem more desirable to simply discard the MLST approach and move on to techniques that better leverage genome-scale data sets. The answer to this question has to do with the vast amount of critically important legacy data that have been generated by the application of MLST schemes to scores of bacterial pathogens over the years. The most widely used MLST scheme database – PubMLST[1] – currently hosts MLST schemes for 99 species (or

[1] https://pubmlst.org/databases/

genera) of bacterial pathogens along with 10 eukaryotic (fungal) pathogens, bacteriophages and plasmids. These schemes cover many tens of thousands of distinct allelic sequences and have been widely applied in hundreds of molecular epidemiology studies around the world, including routine surveillance and outbreak investigations. Together, these data and results represent a wealth of information relating bacterial genome sequence variation to determinants of infectious disease. As such, it will remain critically important to continue characterizing bacterial isolates with respect to their MLST sequence types. Of course, with whole genome sequences in hand, it will also be possible to apply one or more of the new larger-scale typing schemes to the same data sets used to generate MLST sequence types. These two approaches are by no means mutually exclusive.

The remaining importance of MLST in the post-genomic era, combined with the fact that it is now faster and cheaper to sequence whole genomes using NGS platforms than to Sanger sequence MLST amplicons, necessitates the development and application of computational techniques for MLST analysis using NGS datasets. Indeed, there has been a substantial developmental effort for genome-enabled MLST software over the last eight years. As of this writing, there are at least 13 different genome-based computational methods for MLST analysis (Table 1). Our own group developed the program stringMLST, which uses a distinct *k*-mer based approach for genome-enabled MLST to yield extremely rapid and 100% accurate MLST sequence types directly from NGS read data. *k*-mers are sequence substrings, or words, of length *k*. This alignment-free *k*-mer based approach represents a substantial technological advance for computational methods for genome-enabled MLST, which other groups have recently extended.

Table 1. List of Alignment-based and alignment-free methods for multilocus sequence typing.

Computational Tool	Description	Input Data Type	User Interface	Website	Release Year	Reference
Alignment-based method algorithms that utilize de novo assembly, genome mapping and/or sequence alignment						
BIGSdb	Database and analytical platform designed for microbial loci-based typing schemes. Open-source, freeware, locally installable; base platform for PubMLST website; utilizes BLAST	Genome, gene sequences	Web/GUI	https://pubmlst.org/software/database/bigsdb	2010	(Jolley and Maiden, 2010)
MLSTcheck	Automated, scalable command line tool for determining MLST from genome sequences; utilizes BLAST	Genome sequences	CLI	https://www.sanger.ac.uk/science/tools/mlstcheck	2016	(Page, et al., 2016)
MLSTar	R based package to determining MLST from genome sequences; utilizes BLAST	Genome sequences	CLI	https://github.com/iferres/MLSTar	2018	(Ferres and Iraola, 2018)
chewBBACA	Comprehensive pipeline for creation of whole- and core-genome MLST (wgMLST & cgMLST) as well as determining wgMLST/cgMLST from genome sequences using BLAST Score Ratio (BSR)	Genome sequences	CLI	https://github.com/B-UMMI/chewBBACA	2018	(Silva, et al., 2018)
DTU CGE MLST 2.0	Web-based application for performing MLST analysis; utilizes *de novo* assembly and BLAST for MLST	Genome sequences; NGS reads	Web/GUI	https://cge.cbs.dtu.dk/services/MLST	v1: 2012 v2: 2018	(Larsen, et al., 2012)
SRST/SRST2	Read-to-genome mapping based application for performing MLST from NGS read data	NGS reads	CLI	https://katholt.github.io/srst2/	v1: 2012 v2: 2014	(Inouye, et al., 2014)
MOST	Modification of SRST2 for MLST analysis and *Salmonella* serotyping from NGS reads	NGS reads	CLI	https://github.com/phe-bioinformatics/MOST	2016	(Tewolde, et al., 2016)
ARIBA	Pipeline that performs read-to-gene mapping followed by targeted assembly	NGS reads	CLI	https://github.com/sanger-pathogens/ariba	2017	(Hunt, et al., 2017)
Kestrel	Novel algorithm that uses *k*-mers and dynamic programming based local alignment to perform MLST	NGS reads	CLI	https://github.com/paudano/kestrel	2017	(Audano, et al., 2018)
Alignment-free algorithms that do not utilize assembly or alignment based techniques						
stringMLST	Loci-based typing using *k*-mer counting and hash tables	NGS reads	CLI	https://github.com/jordanlab/stringMLST/	2017	(Gupta, et al., 2017)
STing	Computationally efficient implementation of stringMLST; utilizes *k*-mer frequencies and enhanced suffix arrays	NGS reads	CLI	https://github.com/jordanlab/STing	2019	(Espitia-Navarro, 2019; Espitia, et al., 2017)
MentaLiST	Loci-based typing using *k*-mer counting followed by colored de Bruijn graph construction	NGS reads	CLI	https://github.com/WGS-TB/MentaLiST	2018	(Feijao, et al., 2018)
Krocus	Loci-based typing from long read sequencing data; utilizes *k*-mer counting	Long read sequences	CLI	https://github.com/andrewjpage/krocus	2018	(Page and Keane, 2018)

Genome sequence based approaches for MLST can be broadly classified into two groups – (i) classic alignment-based methods that use genome assembly and/or read mapping, and (ii) newer alignment-free approaches that utilize *k*-mers to derive sequence types directly from NGS read data (Figure 3).

1.2.1 Alignment-based computational methods

Alignment-based methods for MLST, or other locus-based typing schemes, entail the comparison between isolate allele sequences and typing scheme databases using sequence similarity searches (Figure 3). A number of these approaches require an assembly step in order to work with short read data generated by NGS platforms. Once the NGS read data are assembled into longer contiguous (contig) sequences, they are compared to allele and profile databases to generate sequence types. Examples of this kind of typing software include BIGSdb (Jolley and Maiden, 2010), MLSTcheck (Page, et al., 2016), and MLSTar (Ferres and Iraola, 2018). Genome assembly is computationally expensive, in terms of both CPU time and memory, and it can require substantial bioinformatics expertise to generate reliable results. As such, assembly represents a major bottleneck for genome-enabled molecular typing studies, and these approaches do not scale well when hundreds of isolates need to be characterized. Assembly-based methods are also difficult to implement for larger scale locus-based typing schemes that employ hundreds or thousands of genome-wide loci.

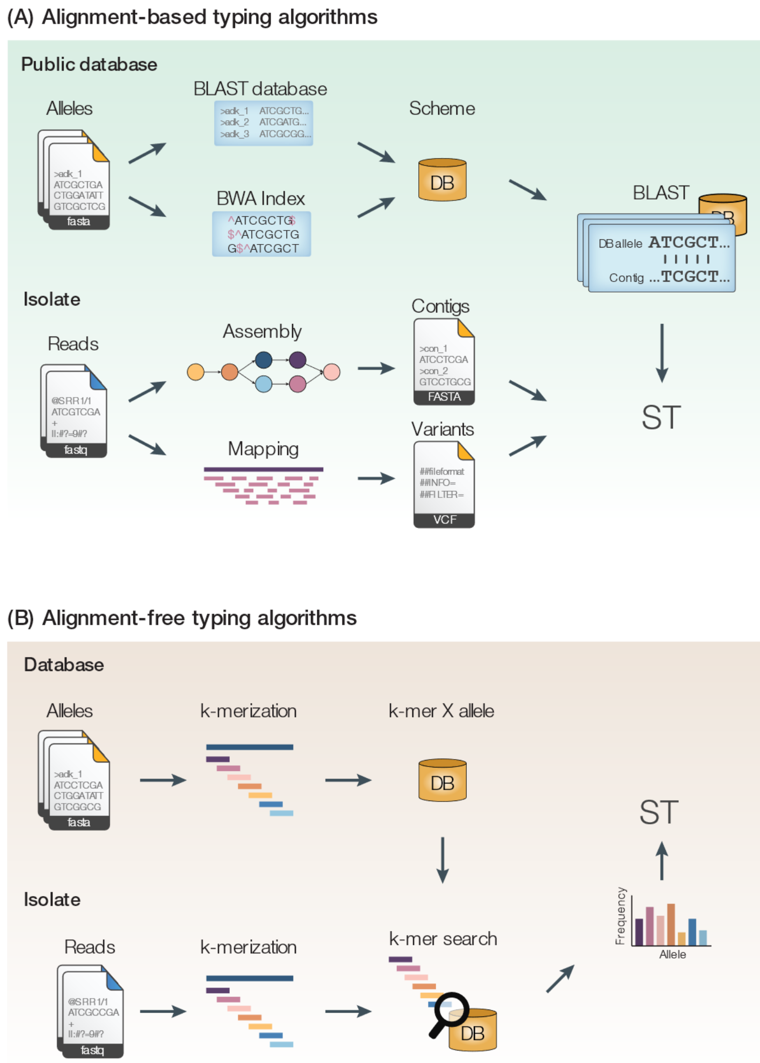


Figure 3. Schematic comparison of alignment-based and alignment-free algorithms for sequence typing. The figure provides the general overview of the two dominant paradigms for performing multilocus sequence typing from whole genome sequence read datasets. Both methods utilize a database of allele sequences for each locus in the scheme and an allelic profile table that contains the mapping of allele numbers to a sequence type. (A) Alignment-based typing algorithms can be further subcategorized into assembly-based and mapping-based. Assembly-based algorithms make use of de novo genome assembly followed by sequence similarity searching algorithms such as BLAST. Mapping-based algorithms map the read sequences to either a reference genome or loci sequences, followed by variant identification. (B) Alignment-free algorithms utilize exact matching of substrings, also known as *k*-mers, between NGS reads and allele sequences in the database to identify the sequence type. Exact substring matching is computationally faster than genome assembly or sequence alignment, and these algorithms gain further speed by comparing only a small fraction of the input read dataset and discarding all non-informative reads.

Another class of algorithms for bacterial typing with NGS data uses short read mapping to reference sequences as a more computationally tractable alternative to assembly-based methods. These methods can still be considered as alignment-based, since they rely on read-to-genome alignments; nevertheless, they are substantially more efficient compared to assembly-based methods. The Center for Genomic Epidemiology[2] provides a genome-based web platform for MLST, which previously implemented an assembly based approach and has since evolved to use read mapping for allele calling (Larsen, et al., 2012). The first program designed specifically to do NGS-based bacterial typing via read mapping was SRST (Inouye, et al., 2014), which was subsequently modified by the same group to develop SRST2 and another group to develop the program MOST for *Salmonella* serotyping (Tewolde, et al., 2016). More recently, the program ARIBA implemented a hybrid approach that uses read mapping to clusters of related alleles followed by constrained assembly of reads that map to specific clusters (Hunt, et al., 2017).

1.2.2 Alignment-free computational methods

The development of alignment-free methods for genome-based molecular typing with NGS data was a major breakthrough that provided substantial increases in speed and efficiency compared to existing assembly or read mapping approaches. As the name implies, these methods proceed directly from raw NGS sequence read data – without any quality control, alignment, or assembly steps – to call alleles and sequence types (Figure 3). The program stringMLST, developed by our group, was the first program of this kind

[2] http://www.genomicepidemiology.org/

designed for bacterial typing directly from NGS data (Gupta, et al., 2017). stringMLST was designed and implemented to provide a turn-key solution of bacterial typing from genome sequence data, with minimal requirements for computational capacity or bioinformatics expertise.

The stringMLST algorithm relies on the use of *k*-mer frequencies and hash tables for characterizing the sequence types of bacterial isolates directly from genome sequence read data. In order to type bacterial isolates from any given species, stringMLST requires a database built from the alleles of the species-specific typing scheme. To construct the typing scheme database, stringMLST generates all possible *k*-mers from each allele sequence in the scheme and stores them in a hash table that associates each *k*-mer with all of the alleles in which it can be found. To characterize an isolate sample, the stringMLST algorithm performs three steps: (i) filtering, (ii) *k*-mer counting, and (iii) reporting. For the filtering step, the algorithm discards a read if the *k*-mer located in the middle of the read sequence is not present in the allele *k*-mer database. This heuristic step provides the bulk of the speed and efficiency to the stringMLST algorithm by passing over reads that correspond to genomic regions not covered by the typing scheme. Since this genomic fraction corresponds to the vast majority of the genome sequence for MLST schemes, only a tiny fraction of the reads needs to be fully processed by the algorithm. For the *k*-mer counting step, if the middle *k*-mer is found in the allele database, then stringMLST generates all possible *k*-mers from the read sequence. The algorithm then searches the read *k*-mers against all *k*-mers in the database and updates a table of *k*-mer frequencies for each associated allele. Steps (i) and (ii) are repeated until all of the reads are processed. For the final reporting step, the algorithm then reports the alleles with the maximum *k*-mer frequency for all loci in the typing scheme, thereby generating an allelic profile and calling the corresponding sequence type.

Compared with existing genome sequence-based typing tools that utilize alignment and/or the assembly, the stringMLST approach is far more efficient and at least as accurate for characterizing bacterial isolates. As reported in Gupta et al. (Gupta, et al., 2017), stringMLST was the only tool able to correctly type each of 40 NGS samples from four different bacterial species (*Campylobacter jejuni*, *Chlamydia trachomatis*, *N. meningitidis*, and *Streptococcus pneumoniae*). It was up to 65x faster than other programs used to process the same datasets, showing an average of 45 seconds to process each sample read file. In the same study, stringMLST correctly predicted the sequence type for 99.8% of 1,002 isolates of *N. meningitidis* requiring an average of 40.7 seconds and 0.67 MB of RAM to type each sample read file. Page et al. (Page, et al., 2017) performed an independent comparison of eight different programs for genome-based MLST, including stringMLST as the only application on the category of alignment-free based methods. In addition to evaluating the accuracy of the tools on NGS data from past outbreaks, they evaluated the impact of sequencing depth and sample contamination on typing speed and accuracy using simulated data. Consistent with our own results, stringMLST was found to be the fastest algorithm by far and also required substantially less computational resources than any of the other programs. In addition, stringMLST proved to be 100% accurate for bacterial typing on outbreak data, comparable to slower and more cumbersome tools that rely on sequence alignment and/or assembly. It is also worth noting that stringMLST does not require any read pre-processing or quality control, making it far easier to use than the other tools and ideally suited for deployment in public health laboratories or in the field.

Despite the superior performance of stringMLST for genome-based MLST, it does suffer from scaling issues when applied to larger-scale typing schemes. We cover these issues, and how we are addressing them, in the subsequent sections on genome-scale typing schemes.

Several other groups have introduced *k*-mer based typing methods since the development of stringMLST. For example, the program Kestrel (Audano, et al., 2018) uses a hybrid approach that combines *k*-mer analysis with dynamic programming based local alignment to call MLST alleles and sequence types. However, this approach is far slower and less efficient than the *k*-mer only method used by stringMLST, which is 28x faster and requires an average of ~60% of the RAM compared to Kestrel. This performance difference is likely due to the Kestrel algorithm's reliance on the exhaustive dynamic programming step. The program MentaLiST (Feijao, et al., 2018) extends the stringMLST approach of using *k*-mer frequencies and hash tables, by constructing a colored de Bruijn graph for each allele of the typing scheme. With this addition, MentaLiST selects a subset of *k*-mers that embodies the variation present in the alleles of the typing scheme, resulting in a substantial reduction in the size of the allele database. This database compression allows for substantial improvement of the computational performance on larger typing schemes that utilize hundreds or even thousands of loci genome-wide. We cover the computational challenges and opportunities entailed by these so-called superMLST schemes in the following section. Yet another example of new *k*-mer based typing software is Krocus (Page and Keane, 2018), designed for typing from uncorrected long-read sequence data. A problem with these kinds of data is that the current long-read sequencing technologies (Pacific Biosciences and Oxford Nanopore) exhibit high error rates.

However, base errors tend to be uniformly distributed, a characteristic exploited by the Krocus developers to circumvent the high error rate problem. Perhaps the most attractive feature of Krocus is that it can type isolates in real-time by taking batches of long-reads produced by sequencers that support continuous sequence streaming like those developed by Oxford Nanopore Technologies.

1.3 Genome-enabled bacterial typing schemes

We previously described why whole genome sequence data are still used for small-scale locus-based typing schemes such as MLST, owing to a combination of the low cost and ease of genome sequencing coupled with the epidemiological importance of MLST legacy data. Nevertheless, the ever increasing availability of numerous whole genome sequences from bacterial pathogens (Figure 2) provides both challenges and opportunities for the development of novel, large-scale typing schemes, which leverage the analysis of genome-wide variation data. Genome-scale bacterial typing schemes can be broadly categorized as (i) locus-based schemes, or (ii) single nucleotide variant (SNV)-based schemes (Table 2). Locus-based typing schemes are direct extensions of MLST that rely on the analysis of hundreds or thousands of loci genome-wide, as opposed to the handful of loci used by MLST schemes. For example, core genome MLST (cgMLST) schemes utilize all of the loci that correspond to the core genome with all genes shared among a set of isolates (*i.e.* the intersection of genes in a set of genomes). Whole-genome MLST (wgMLST) schemes are even larger-scale and use all of the genes (*i.e.* the union) found in a set of genomes; this approach includes both the core-genome and the accessory-genome.

These large-scale loci-based bacterial typing schemes provide substantially more resolution than traditional MLST schemes.

Table 2. Comparison of locus-based and single nucleotide variant (SNV)-based typing techniques for bacterial typing.

	Locus-based Typing	**Single Nucleotide Variant (SNV) Typing**
Advantages	• Ideal for microbial genome analysis • Allows for comparisons between different studies/outbreaks • Each isolate can be easily computationally represented in a defined space • Availability of several online, publicly accessible resources (tools and large databases) • Standardized pipelines are available • Can be configured to analyze core and accessory genome • Phylogeny reconstruction methods are simpler in nature (UPGMA, eBURST)	• Ideal for complex Eukaryotic genome analysis, such as human • High level of discrimination power; allows inspection of every single nucleotide change across the genome • Works well if a reference genome is standardized and internationally used • Cn be detected using both sequencing and real-time PCR-based methods • Diagnostic SNVs exists for the subtyping agents
Disadvantages	• Requires a curated database of alleles and profile definitions • Loci based schemes are often restricted to genic regions and does not capture variation in intergenic (or intronic) regions • Captures gene presence/absence but fails to capture other large structural variations *e.g.*, duplications and rearrangements	• Comparison between different studies/outbreaks is limited due to differences in reference genome • Requires an evolutionarily close reference genome, preferably finished • Mostly captures the core genome; misses variations in accessory genome • SNV calls are dependent on filtering criterion used • Does not capture large structural variation events, viz., insertion/deletion (indels), duplications and rearrangements • Computational storage grows exponentially as SNV data typically involves representing all sites across the genome • Commonly used phylogenetic methods are computationally intensive (Neighbor-Joining, Maximum Likelihood) • Fails to capture high horizontal gene transfer (HGT)

In principle, SNV-based approaches to genome analysis provide even more resolution for the delineation of bacterial lineages than the largest-scale loci-based schemes, because there are far more possible single base differences among genomes than the possible number of differences among loci. As such, SNV-based schemes should be able to differentiate extremely closely related strains, down to 1 bp difference in principle. This feature makes SNV-based approaches better for studies that require extreme levels of resolution, such as contact tracing studies that seek to characterize the exact origins of bacterial outbreaks and their spread among patients (Stucki, et al., 2015). A classic example of this approach is the single base pair resolution typing of *Vibrio cholerae* strains from the 2010 outbreak in Haiti, which ultimately pointed to United Nations peacekeepers from Nepal as the source of the outbreak (Katz, et al., 2013).

Nevertheless, there are a number of reasons why locus-based typing schemes are still widely employed for bacterial typing in the post-genomic era. Perhaps most importantly, locus-based schemes are portable and more reproducible than SNV-based schemes. Since they rely on a pre-defined set of loci, and the accompanying allele databases, locus-based schemes generate results that can be directly compared among laboratories and among different studies. SNV-based schemes rely on the use of one or more reference sequences for variant calling and are thereby more difficult to standardize among groups. The use of reference sequences for variant calling with SNV-based schemes can also lead to a loss of information with respect to accessory genes, which are often important determinants of virulence for bacterial pathogens. Locus-based schemes, on the other hand, can readily accommodate important accessory genes via presence/absence calls for those loci. SNV-based typing is not suitable for organisms that undergo high rate of horizontal gene transfer as the number of nucleotide differences does not correlate well with time. Additional details on the relative strengths and weaknesses of locus-based versus SNV-based typing schemes can be found in Table 2. Given the continued importance of locus-based typing schemes for genome-enabled bacterial typing, we focus on the computational approaches used for these kinds of schemes in the following sections.

1.3.1 Computational approaches to large-scale typing schemes

As with MLST, large-scale bacterial typing schemes that leverage genome-wide data sets can be computationally implemented using traditional alignment/assembly-based methods or with the newer *k*-mer based approaches. However, it is becoming increasingly apparent that the traditional methods lack the computational speed and efficiency needed to implement such schemes for rapid bacterial typing. For example, approaches that use *de novo* assembly followed by BLAST can take upwards of 12 hours for each isolate for cgMLST and/or wgMLST schemes, which require the analysis of thousands of loci per isolate. As such, the traditional methods will become increasingly irrelevant for epidemiological studies that need to type scores, hundreds, or even thousands of isolates. For this reason, we focus here on the latest developments in the computational approaches for large-scale bacterial typing schemes.

We previously discussed how the application of the first *k*-mer based approaches for bacterial typing in the stringMLST algorithm resulted in orders of magnitude speed-up for MLST without any loss of accuracy. However, the stringMLST algorithm did not scale well to large-scale typing schemes like cgMLST. When stringMLST was applied to schemes of this kind, it did not compute any faster than alignment/assembly-based approaches and required an unrealistically large amount of memory to run. This performance was because the underlying hash-table data structure used for the allele *k*-mer database are not optimally suited for large-scale typing schemes, since it entails the storage of all existing *k*-mers for thousands of loci. As previously discussed, the more recently developed program MentaLiST addressed this challenge by using a de Bruijn graph to substantially compress the allele *k*-mer database while also providing for enhanced searching of the database. This revised data structure provides for robust – rapid and accurate – bacterial typing using large-scale typing schemes directly from NGS read data. Our own group is currently developing the algorithm STing (as a successor to stringMLST) that employs a more efficient data structure, thereby allowing for genome-based typing with large-scale schemes.

STing is being developed and implemented for both bacterial typing and gene detection directly from unprocessed NGS read data (Espitia, et al., 2017). The STing algorithm stores the allele *k*-mer databases for large-scale typing schemes using an enhanced suffix array data structure as opposed to the simpler hash table used by stringMLST. The suffix array provides for a substantially compressed representation of the allele *k*-mer database as well as rapid search capability along the array. STing has been applied to MLST, cgMLST, and wgMLST schemes for a wide variety of bacterial pathogens. It can also be used for automated gene detection directly from read sequences, and this utility is currently being validated in the context of antimicrobial resistance genes and virulence factors (*e.g.* Shiga toxin). Preliminary results on the performance of STing are very promising, and a more detailed description of both the algorithm and its accuracy is currently in preparation.

1.3.2 Community adoption of genome-based bacterial typing

As we have mentioned several times, the genome revolution provides both amazing opportunities and profound challenges to the public health community. In principle, genome sequence data provide for unprecedented levels of resolution for bacterial typing, while also generating abundant material for the discovery of the genetic determinants of antibiotic resistance or virulence. Nevertheless, there are substantial technical hurdles that need to be overcome in order to ensure that the community can fully adopt genome-enabled approaches to molecular epidemiology along with the new bioinformatics techniques that they necessitate.

One key feature of early sequence-based bacterial typing schemes – MLST in particular – was portability in terms of the ability to broadly share uniformly comprehensible typing results among member laboratories distributed among surveillance networks. Portability refers to both the typing techniques, which should be standardized so that they can be carried out in any laboratory, and the typing results, which should have the same representation irrespective of where the results are generated. MLST is ideally suited for portability since it relies on a shared set of loci (allele) sequence definitions and produces granular and static sequence types from the typing scheme's allelic profiles. Larger-scale typing schemes face a number of challenges in order to ensure that they both (i) remain completely portable and (ii) allow for comparison with the results of previous generation typing techniques.

The challenge to portability for genome-scale typing schemes is directly related to the scale of these schemes, which can cover hundreds or thousands of loci genome-wide. The large scale of these schemes necessitates a highly coordinated effort to standardize the loci (allele) definitions that underlie the schemes and entails far more complicated allelic databases than is the case for MLST schemes, which typically utilize 7-9 loci. With respect to loci definitions, there needs to be an agreement concerning exactly which loci are used for any scheme and which part (*i.e.* sequence fragment) of each locus is used for typing.

This aspect is relatively straightforward for schemes with a few loci but is substantially more complex when hundreds or thousands of loci are used. Furthermore, since genome-scale typing schemes are being independently developed in multiple public health laboratories around the world, numerous different versions of the same typing scheme can end up being used. With respect to allelic databases, despite the fact that thousands of bacterial pathogen genome sequences have already been characterized, allelic and profile databases for larger schemes are either incomplete or do not yet exist. A coordinated effort by the public health community will be needed to address these issues and ensure that genome-enabled typing schemes remain standard and portable. This process needs to happen soon, since it will be difficult for individual laboratories, or particular surveillance networks, to change their typing schemes once they are developed and implemented.

Another critical issue for genome-enabled typing schemes will be the ability to maintain some connection to the vast amount of historical information contained in results generated from smaller scale legacy typing schemes. In other words, genome-scale typing schemes should be backward compatible, to whatever extent possible, with previous typing schemes such as MLST or even the non-sequence-based pulsed-field gel electrophoresis (PFGE) typing scheme. Public health laboratories will need to dedicate a substantial amount of bioinformatics expertise and effort to map the results of genome-scale typing schemes to the results of legacy typing schemes. An illustrative example of this challenge is the U.S. Centers for Disease Control and Prevention (CDC) PulseNet surveillance network[3]. PulseNet was established in 1996 as a network of public health laboratories around the US dedicated to surveillance and outbreak detection for food and waterborne

[3] https://www.cdc.gov/pulsenet/

illness caused by a prioritized set of bacterial pathogens. PulseNet laboratories use a restriction enzyme based technique to digest genomes of bacterial pathogen isolates. PFGE generates characteristic DNA fingerprints of the digested genomes, which are captured as distinct banding patterns on a gel. The implementation of PFGE across the PulseNet surveillance network allowed for the discovery of clusters of disease that corresponded to outbreaks, thereby leading to better coordinated and more rapid responses to such public health threats. PulseNet's use of the relatively low resolution and clearly outdated PFGE technique is expected to be phased out starting in 2019, after which time reliance will be exclusively on the far higher resolution genome-enabled typing schemes. Nevertheless, given the amount of invaluable epidemiological information that is tied to specific PFGE patterns, it will be critically important to be able to relate the results of genome-scale typing schemes to previously characterized patterns. Accordingly, CDC scientists are working to develop approaches for the probabilistic association of PFGE patterns and genome sequence variation, and our own laboratory is involved in this effort via a collaboration with the CDC's Enteric Diseases Laboratory Branch within the Division of Foodborne, Waterborne, and Environmental Diseases (DFWED).

The challenges for genome-enabled typing schemes outlined above – relating to uniform data standards, typing scheme portability, and backward compatibility – also suggest a pressing need for shared analytical platforms that can be deployed in public health laboratories around the world. Generating whole genome sequence data is now rapid, cost effective, and highly standardized. Accordingly, the rate-limiting step for genome-enabled bacterial typing corresponds to the suite of computational analysis tools and methods that need to be used to handle and interpret the massive volumes of data generated by NGS platforms. Here, we are considering mainly the software challenges entailed by the use of NGS data for bacterial typing, but there are also substantial hardware issues that need to be addressed. The sheer volume of data alone poses a fundamental challenge with respect to both computational storage and processing capacity. It is not realistic to expect that all public health laboratories will be able to address these joint challenges via the deployment of local computational capacity. In fact, we are closely reaching the point where it will cost less to sequence bacterial genomes than to store the resulting sequence data for an extended period of time. Similarly, the computational processing power needed to handle hundreds or thousands of genome sequences of bacterial isolates is likely out of reach for all but the most well-funded public health laboratories.

Cloud computing environments, whereby computational storage and processing are provisioned as services that are accessed remotely over the internet, offer an attractive alternative to the deployment of local computational capacity for bacterial genome analysis. One of the most compelling features of cloud computing is the flexibility entailed by the on-demand model whereby investigators only make use of the amount of computational capacity that they need at any given moment. This relates to both processing power, in terms of the number and architecture of compute cores that can be accessed for any given analysis, and the elastic nature of cloud data storage capacity, with different models of data access for short term and longer term storage. Over the last five years, there has been a concerted effort to deploy computational genomics algorithms and pipelines across a variety of cloud computing platforms. In Table 3, we show examples of cloud computing resources in support of bacterial genome analysis with respect to both specific bioinformatics software packages as well as integrated bioinformatics platforms. The integrated platforms allow users to utilize existing bioinformatics analysis pipelines and/or build their own custom pipelines, which employ multiple applications to execute an entire workflow.

Table 3. Examples of integrated bioinformatics cloud computing software and platforms for microbial genome analysis.

Resource	Description	Website	Reference
Illumina BaseSpace	Illumina's platform for subscription-based bioinformatics data analysis	https://basespace.illumina.com/	-
RAST	Automated microbial genome annotation pipeline	http://rast.nmpdr.org/	(Aziz, et al., 2008)
Galaxy	Open-source, free-to-use software for a variety of bioinformatics data analyses. Cloud support through Amazon Web Services, CloudMan, Globus Genomics	https://usegalaxy.org/	(Afgan, et al., 2018)
CloudBioLinux	Community driven, cloud-based bioinformatics platform	http://cloudbiolinux.org/	(Krampis, et al., 2012)
CLIMB	UK's nationwide bioinformatics/electronic infrastructure designed to support the needs of the microbiology community	http://bryn.climb.ac.uk/	(Connor, et al., 2016)
CloVR	A desktop bioinformatics virtual machine capable of utilizing cloud computing resources	http://clovr.org/	(Angiuoli, et al., 2011)
Nephele	Cloud platform for microbiome data analysis	https://nephele.niaid.nih.gov	(Weber, et al., 2018)

Despite the promise of the cloud computing model for computational genomics, there is currently no standardized cloud computing platform to support genome-enabled bacterial typing. Given the explosion of bacterial genome sequences, coupled with the development of numerous genome-scale typing schemes, we anticipate a pressing need for the cloud deployment of a standardized genome analysis platform in support of genome-enabled bacterial typing in public health laboratories. A shared analytical platform of this kind should consist of (i) a uniform set of bioinformatics analysis tools, (ii) a shared set of standard analysis protocols or pipelines for the use of these tools and (iii) a set of well-defined data models that covers both input and output standards for the bioinformatics tools as well as the loci (allele) databases that underlie bacterial typing for multiple schemes across multiple species. This platform should also include mechanisms for storing primary NGS data as well as secondary data (results) generated by the analytical platform along with transparent means for sharing data and communicating results among public health laboratories. Finally, the use of a unified approach to collect and distribute epidemiological meta-data associated with bacterial isolates characterized as a part of routine surveillance and outbreak investigations will also be a critical component for such a platform.

We envision that the integrated cloud computing service and the standardized bacterial genome analysis platform described above could be unified into a national or global surveillance network with constituent public health laboratories as nodes that are capable of both rapidly typing bacterial isolates and widely sharing the results with other laboratory nodes around the world. To our knowledge, no such integrated platform currently exists, and perhaps even more disconcerting, there is a real possibility that genome-enabled approaches to bacterial typing will ultimately hamper efforts to share bacterial typing results among different laboratories. In particular, if different public health laboratories continue to independently develop their own genome-scale typing schemes, it will become increasing difficult, if not impossible, to meaningfully compare results among laboratories. Obviously, such an outcome should be avoided at all costs; it would be truly unfortunate if the increased resolution afforded by genome-scale typing schemes paradoxically leads to less resolution on the public health challenges to which these schemes are ultimately addressed. Ensuring that such a scenario does not come to pass will require an ongoing effort towards the development, standardization, and sharing of computational approaches to, and platforms for, genome-enabled bacterial typing.

CHAPTER 2. STING: ACCURATE AND ULTRAFAST GENOMIC PROFILING WITH EXACT SEQUENCE MATCHES

2.1 Abstract

Genome-enabled approaches to molecular epidemiology have become essential to public health agencies and the microbial research community. We developed the algorithm STing to provide turn-key solutions for molecular typing and gene detection directly from next-generation sequence data of microbial pathogens. Our implementation of STing uses an innovative *k*-mer search strategy that eliminates the computational overhead associated with the time-consuming steps of quality control, assembly, and alignment required by more traditional methods. We compared STing to six of the most widely used programs for genome-based molecular typing and demonstrate its ease of use, accuracy, speed, and efficiency. STing shows superior accuracy and performance for standard multilocus sequence typing schemes, along with larger genome-scale typing schemes, and it enables rapid automated detection of antimicrobial resistance and virulence factor genes. We hope that the adoption of STing will help to democratize microbial genomics and thereby maximize its benefit for public health.

2.2 Introduction

Molecular typing entails the identification of distinct evolutionary lineages (*i.e.*, types) within species of bacterial pathogens; it is an essential element of both outbreak investigation and routine infectious disease surveillance (Espitia-Navarro, 2019; Maiden, et al., 2013). Multilocus sequence typing (MLST) was developed as the first sequence-based approach to molecular typing in 1998 (Maiden, et al., 1998). Initially, MLST schemes relied on Sanger sequencing of PCR amplicons from fragments of 7-9 housekeeping genes spread throughout the genome. While this approach truly revolutionized molecular epidemiology, it is time consuming and costly compared to current next generation sequencing (NGS) methods. Nevertheless, MLST remains widely used for molecular typing, particularly in light of valuable legacy data relating sequence types (STs) to epidemiological information.

Public health agencies increasingly couple NGS characterization of microbial genomes with downstream bioinformatics analysis methods to perform molecular typing. The overhead associated with the bioinformatics methods that are used for this purpose, in terms of both the required human expertise and computational resources, represents a critical bottleneck that continues to limit the potential impact of microbial genomics on public health. This is particularly true for local public health agency laboratories, which are typically staffed with microbiologists who may not have substantial bioinformatics expertise or ready access to high-performance computational resources. In light of this ongoing challenge, our group is working to develop turn-key solutions to genome-enabled molecular epidemiology, including both molecular typing and the detection of critical antimicrobial resistance (AMR) and virulence factor (VF) genes. Methods of this kind must be easy to use, computationally efficient, fast, and most importantly, highly accurate.

We previously developed stringMLST as an alternative approach to genome-enabled molecular typing of bacterial pathogens (Gupta, et al., 2017). stringMLST relied on *k*-mer matching between NGS sequence reads and a database of MLST allele sequences, thereby eliminating the need for the sequence quality control, genome assembly, and alignment steps that the first generation of genome-enabled typing algorithms used. It proved to be accurate and fast for traditional MLST schemes, but it did not scale well to the larger genome-scale typing schemes, such as ribosomal MLST (rMLST) or core-genome MLST (cgMLST), which are increasingly used in molecular epidemiology (Jolley, et al., 2012; Maiden, et al., 2013). Here, we present our new approach to this problem – STing. The STing algorithm is distinguished from its predecessor in several important ways: the efficiency of its code base, the underlying data structure that is uses, and the scope of its applications. These innovations provide for superior accuracy and performance compared to both stringMLST and other widely used programs for genome-enabled molecular typing. Below, we provide a high level overview of the STing algorithm, details of which can be found in the Methods section, and we report on its use across several typing schemes and for automated gene detection in the Results section.

The STing algorithm breaks down (*k*-merizes) NGS reads into *k*-mers and then compares read *k*-mers against an indexed reference sequence database (Figure 4). The speed and efficiency of the algorithm are derived from the nature of the *k*-mer search strategy that it uses along with the structure of the reference sequence database. For each individual read, a single central *k*-mer is initially compared against the sequence database. Reads are only fully *k*-merized if there is an initial match between the central *k*-mer and the database. If there is no match, which occurs for the vast majority of reads, the read is discarded. Reads are only fully *k*-merized if there is an initial match between the central *k*-mer and the database. If there is no match, which occurs for the vast majority of reads, the read is discarded. This results in substantial savings in terms of both the number of reads that need to be *k*-merized and the number of database search steps.

The reference sequence database is indexed as an enhanced suffix array (ESA) (Abouelhoda, et al., 2004); this enables the efficient representation of entire sequences, as opposed to other *k*-mer based methods that employ *k*-merized sequences in hash tables. The ESA data structure allows for a single sequence index, independent of *k*-mer size, whereas the hash table approach necessitates independent indices for each *k*-mer size. Finally, the ESA data structure facilitates rapid exact *k*-mer matches between input reads and the indexed database. STing can be run in two modes – sequence typing or gene detection – and typing can be run in fast or sensitive modes.

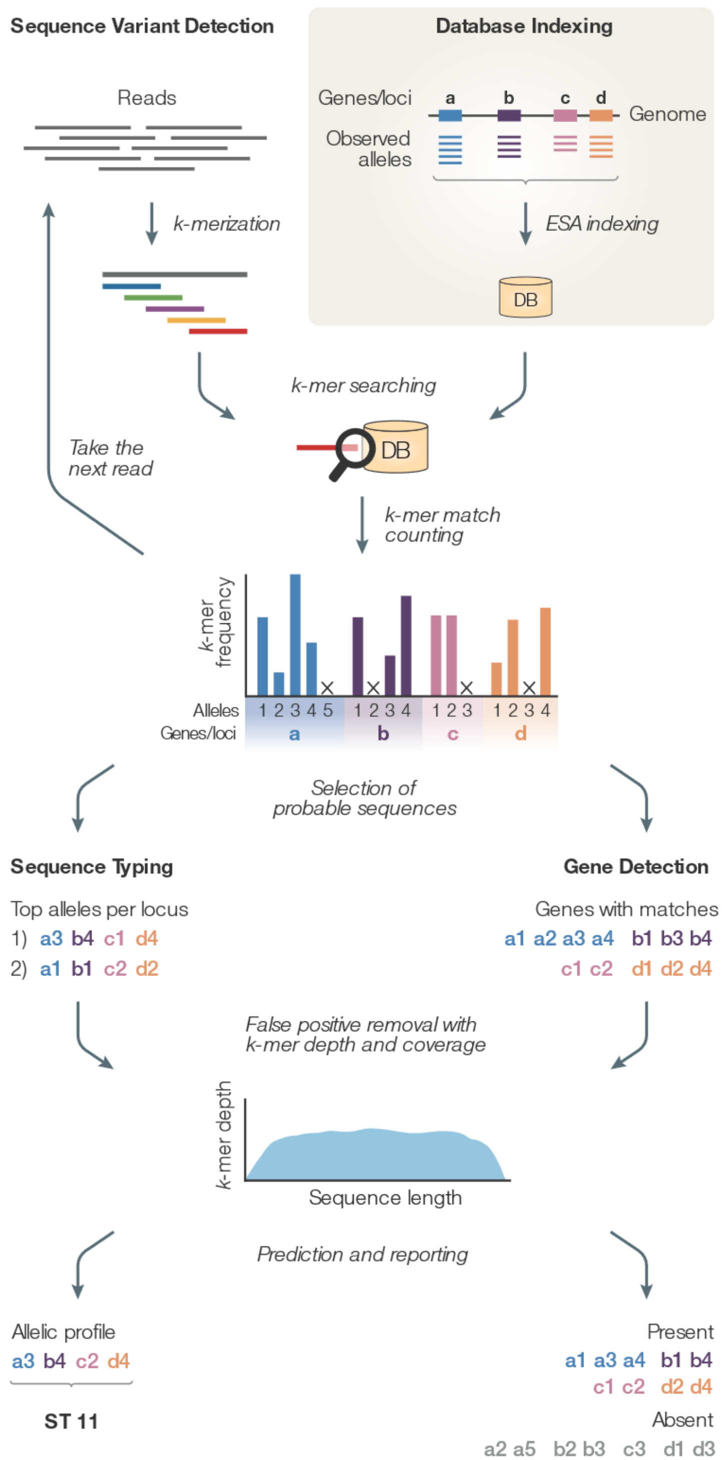


Figure 4. Schematic representation of the STing algorithm. The STing algorithm comprises two main phases: Database indexing (shaded box) – user supplied reference sequences (allele or gene sequences) are transformed into an ESA index for rapid *k*-mer search during the sequence variant detection phase; and Sequence variant detection – reads are *k*-merized and each *k*-mer is searched within the database. For each match located in

the database, a table of frequencies is maintained for the matched sequence within the database. These frequencies are then utilized to select candidate alleles/genes to be present in the samples analyzed. False positive alleles/genes are filtered out by calculating and analyzing *k*-mer depth and sequence length coverage from the selected candidate sequences. Lastly, predictions of allelic profile and ST, and presence/absence of genes, are made and reported. A more detailed flowchart of the algorithm can be seen in Figure 5.

2.3 Materials and methods

2.3.1 Algorithm overview

Given an input sequence read file from a microbial isolate, STing can accurately identify the specific sequence type (ST), *e.g.*, multilocus sequence type or its variants, for the isolate, and what genes of interest are present in its genome. STing accomplishes these tasks by using an exact *k*-mer matching and frequency counting paradigm. STing is implemented in C++ and utilizes two libraries: the SeqAn library (Reinert, et al., 2017) for the ESA data structure, and the gzstream library[4] for working with gz files. Additionally, STing is prepackaged with an R script for visualization of the results and a Python script for downloading database sequences from PubMLST. The ESA data structure is used for *k*-mer look-up and comparison purposes. ESAs are a lexicographically sorted array-based data structure, which represent space efficient implementation of the Suffix Trees data structure. For a given set of sequences with a total length of *n* base pairs (summation of length of all sequences), an ESA index can be constructed in linear time *O(n)*. ESAs can also be queried for *k*-mer matches (or substring matches) in linear time. Given a *k*-mer of length *k*, we can determine its presence/absence in the database in *O(k)* time and find all of

[4] https://www.cs.unc.edu/Research/compgeom/gzstream/

its z occurrences in $O(k+z)$ time. While Suffix Trees achieve the same time complexity for index construction and k-mer lookup, they take five times more storage space than ESAs. An efficient implementation of a Suffix Tree can use up to 20 bytes per input database character, whereas an equivalent ESA consumes 4 bytes per input database character. Using ESAs for k-mer lookup and comparison allows STing to efficiently scale with large sequence databases. The STing algorithm is divided into three steps: (1) database indexing, (2) sequence typing, and (3) gene detection (Figure 5). Each step is described in the following sections.

2.3.2 Database indexing.

In this step, STing constructs an ESA index that is used during the sequence typing and gene detection modes. For sequence typing, the indexer requires a multi-fasta file with all the observed alleles in a typing scheme, and an additional allelic profile file that contains combinations of allele numbers (also referred to as allelic profiles) uniquely mapped to distinct STs. The indexer constructs two ESA indices, one for the allelic sequences (allele index) and one for the profile definitions (profile index). For gene detection, the indexer requires a multi-fasta file with the gene sequences that are to be screened in the input samples. Then, the indexer constructs a single ESA index of all the gene sequences provided (gene index).

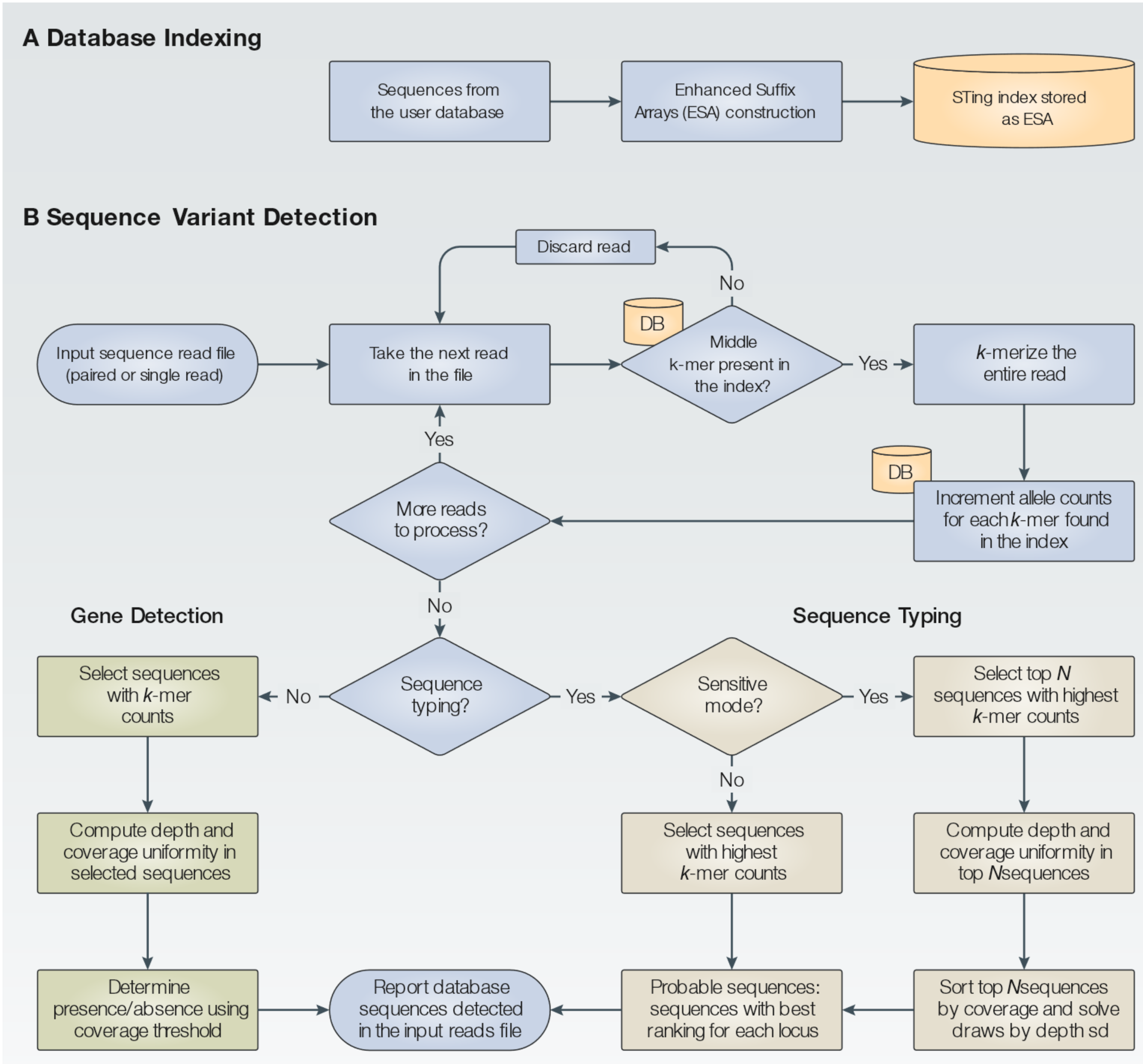


Figure 5. Detailed flowchart of the STing algorithm. The STing algorithm comprises two main phases: (A) Database indexing – user supplied reference sequences (allele or gene sequences) are transformed into an ESA index database for rapid *k*-mer search during the sequence variant detection phase; and (B) Sequence variant detection – The middle *k*-mer of each read is searched within the ESA index. If the middle *k*-mer is not found in the index, the read is discarded; otherwise, the read is passed to the next step. This filtering step allows for massively increased processing speed, since the vast majority of sequence reads in a whole genome sample do not correspond to the loci/genes found in the reference database. Reads that passed the match filter are fully *k*-merized, and each *k*-mer is searched in the ESA index. For each match in the index, a table of frequencies (*k*-mer frequency table) is updated for the matched alleles. For the sequence typing task (tan colored boxes), the top N alleles that have the maximum *k*-mer frequencies of each locus of the typing scheme are selected as candidate sequences to be present in the sample analyzed. If the fast mode is enabled, the value of N is 1, and the alleles for generating the allelic profile are selected based solely in the *k*-mer frequencies. If the sensitive mode is enabled, the

value of *N* is 3 by default (can be configured by the user). Then, for the sensitive mode, the allele coverage and *k*-mer depth is calculated for each of the *N* alleles in each locus. This allows for removing false positive alleles on each locus by identifying pronounced valleys in the *k*-mer distribution. Alleles are called by taking the alleles with maximum length coverage. Ties in length coverage between alleles in the same locus are solved by taking the allele with the minimum *k*-mer depth standard deviation. Allelic profiles are generated with the called alleles and a look up operation of these profiles is performed in the profiles provided by the user to identify the corresponding ST for the sample. Then, the allelic profile and it corresponding ST are reported. For the gene detection task (green colored boxes), the sequences that have at least one *k*-mer match are selected as candidate sequences to be present in the sample. Then, *k*-mer depth and gene coverage is calculated for each selected gene. Genes with a length coverage greater than a threshold provided by the user (75% by default), are considered to be present in the sample analyzed. Finally, presence/absence is reported for each of the selected genes.

2.3.3 Sequence typing.

In this mode, the typer identifies the ST of a given isolate by using a gene-by-gene approach. The typer utility operates in fast or sensitive execution modes. The sequence typing step comprises six algorithmic steps: (1) read filtering, (2) *k*-mer counting, (3) candidate sequence selection, (4) depth and coverage calculation, (5) allele calling and ST prediction, and (6) reporting (Figure 6).

In the read filtering step, the middle *k*-mer of each sequence is searched within the allele index database. If the middle *k*-mer is not found in the allele index, the read is discarded, otherwise the read is passed on to the next step. The size of the *k*-mer is chosen in such a way as to minimalize the possibility that using the middle *k*-mer only results in the loss of useful sequence reads (default *k*=30); users can change the *k*-mer size. In the *k*-mer counting step, the typer *k*-merizes each read that passed the filter matching step, and then searches each *k*-mer from the read against the allele sequence index. For each *k*-mer match in the allele index, the typer increments a *k*-mer counter for the matched alleles/loci.

Once all of the reads are processed, the typer normalizes the *k*-mer frequencies by the length of the corresponding allele.

Algorithm 1: STing Sequence Typing

Input :

Loci $L = \{l_1, l_2, \ldots, l_m\}$, m is the total number of loci.

Reads $R = \{r_1, r_2, \ldots, r_n\}$, n is the total number of reads.

k-mer size $k \leq \min(\text{length}(R))$.

Allele index $\mathcal{A}$, the generalized ESA index of all the alleles A in the typing scheme, where $A = \{a_{(i,j)} \mid 1 \leq i \leq m, 1 \leq j \leq e_i\}$, e_i is the total number of alleles for the locus l_i.

Profile index $\mathcal{P}$, the generalized ESA index of all the allelic profiles P in the typing scheme, where $P = \{p_1, p_2, \ldots, p_q\}$, q is the total number of observed allelic profiles.

Profile table $T = \{t_i \mid 1 \leq i \leq q\}$, where $t_i = (s_i, p_i)$, s_i is the ST associated to the profile p_i.

Output:

Predicted ST and allelic profile $t' = (s_i, p_i)'$ from the read set.

```
procedure SequenceTyping(L, R, k, 𝒜, 𝒫, T)
    for each r ∈ R do                                              ▷ Read processing
        mid_kmer ← GetMidKmer(k, r)                                ▷ (i) Filtering
        if mid_kmer ∉ 𝒜 then
            continue

        freqs, hits ← []                                           ▷ (ii) k-mer counting
        K ← GetAllKmers(k, r)
        for each kmer ∈ K do
            (locus, matched_allele, hit_pos) ← Find(kmer, 𝒜)
            if matched_allele ≠ ∅ then
                freqs[locus][matched_allele]      freqs[locus][matched_allele] + 1
                hits[locus][matched_allele].add(hit_pos)

    norm_freqs ← NormalizeFreqs(freqs, 𝒢)                          ▷ Normalize by gene length

    prob_seqs ← []                                                 ▷ (iii) Selection of probable sequences
    for each locus ∈ L do
        prob_seqs.add(arg max(norm_freqs[locus]))

    (depths, coverages) ← InitDepthsAndCoverages(L, allelic_profile)

    for each allele ∈ allelic_profile do                           ▷ (iv) Depth and coverage calculation
        depths[allele] ← CalculateDepth(hits[allele])
        coverages[allele] ← CalculateCoverage(depths[allele])

    (ST, allelic_profile) ← PredictST(prob_seqs, 𝒫)                ▷ (v) ST prediction

    print ST                                                       ▷ (vi) Reporting
    for each allele ∈ allelic_profile do
        if coverages[allele] < 100 then
            print allele + "*"
        else
            print allele
```

Figure 6. Detailed STing sequence typing algorithm. Input for the sequence typing feature comprises the sequencing reads to process, *k*-mer size, a list of the loci of the sequence tying scheme, and the database elements including the profile table, allele and

profile ESA indices. The output of the algorithm is the predicted allelic profile and its corresponding ST.

In the candidate sequence selection step, the algorithm selects the top *N* alleles that have the maximum normalized *k*-mer frequency for each locus. For the fast execution mode, the default value of *N* is 1, and for the sensitive execution mode is 3 and can be configured by the user.

In the depth and coverage calculation step (only applicable in sensitive mode), the typer reduces the false positives by identifying regions of the candidate alleles that are not covered by any *k*-mer, and identifying any sharp valleys in the *k*-mer depth distribution across the candidate allele. This step calculates the number of *k*-mers that had a match at each base of the top *N* alleles in each locus. To speed-up this calculation, the typer constructs a smaller index consisting of only the top *N* candidate alleles, and parses the subset of reads that passed the initial *k*-mer filter (useful reads). The typer *k*-merizes the useful reads and records the location (base) of each *k*-mer in the matched allele of the smaller index. The algorithm calculates the *k*-mer depth at each base along each allele using the match start positions. The typer then looks for discontinuities in the *k*-mer depth by checking the *k*-mer depth ratio of each adjacent position. The application detects a discontinuity if the ratio is outside the range of $[1/\sqrt{2}, \sqrt{2}]$ and sets the *k*-mer depth as zero for those positions. Finally, the tool calculates the allele coverage as the percentage of allele (*i.e.*, the allele sequence length) that has a non-zero *k*-mer depth.

In the allele calling and ST prediction step, STing generates the allelic profile and predicts the corresponding ST of the sample. For the fast mode, the allelic profile is generated from the candidate sequences selected in the previous step. For the sensitive mode, the allele with the maximum allele coverage for each locus is predicted to be the allele present within the isolate. Here, there are three special cases: (a) in the event that the allele coverage is less than 100%, the detector appends a * character to denote a possible novel allele; (b) in the event of having ties in coverage between alleles, STing calls the allele that has the most uniform *k*-mer coverage by selecting the one with the minimum *k*-mer depth standard deviation; (c) if a locus has no matching *k*-mers, the locus is assumed to be absent and an NA allele is assigned as its call. At this step, all the allele calls have been made and an allelic profile has been generated. A look-up operation is performed in the profile index to identify the ST corresponding to the predicted allelic profile.

Finally, in the reporting step, STing reports the allelic profile, associated ST, and the total number of *k*-mer matches and reads processed, along with optional information about each allele: normalized counts of *k*-mer matches, coverage, and average and per-base *k*-mer depth.

2.3.4 Gene detection.

The algorithm for this mode is a variant of the sequence typing mode and follows the steps described above closely. The gene detection mode differs from the sequence typing mode in how it selects the candidate sequences. This phase can be divide into five conceptual steps: (1) read filtering, (2) *k*-mer counting, (3) candidate sequence selection, (4) depth and coverage calculation, and (5) reporting (Figure 7).

In the *k*-mer filtering step, the detector searches the middle *k*-mer of each read within the gene index. If the *k*-mer fails to match any sequence within the index, the read is discarded, otherwise it is passed on to the next step. In the *k*-mer counting step, the utility proceeds to *k*-merize the read in entirety and searches each *k*-mer in the gene index. A gene-specific *k*-mer match counter is incremented for each *k*-mer that matches the corresponding gene(s). In addition, the detector also records the start position of the *k*-mer in the matching gene(s). In the candidate sequence selection step, STing selects the gene sequences that have at least one *k*-mer match as probable genes present in the sample analyzed. In the depth and coverage calculation step, similar to the sequence typing mode, STing looks for discontinuities in the *k*-mer depth by inspecting the (a) the number of bases not covered by any *k*-mer, and (b) any sharp valleys within the *k*-mer distribution. Next, the tool reports the percent sequence coverage of each gene identified in the sample. Finally, in the reporting step, STing determines the presence/absence of genes with *k*-mer hits. A gene is predicted to be present if its coverage is equal or greater than a user specified threshold (default = 75%). Otherwise, the gene is predicted to be absent in the sample. STing reports the presence (reported as 1)/absence (reported as 0) of each gene with *k*-mer matches, the total number of *k*-mer matches and reads processed, along with optional information about each gene: normalized counts of *k*-mer matches, coverage, and average and per-base *k*-mer depth.

Algorithm 2: STing Gene Detection

Input :

Genes of interest $G = \{g_1, g_2, \ldots, g_m\}$.

Reads $R = \{r_1, r_2, \ldots, r_n\}$, n is the total number of reads.

k-mer size $k \leq \min(\text{length}(R))$.

Coverage threshold θ.

Gene index $\mathcal{G}$, the generalized ESA index of G.

Output:

Presence/absence predictions $P = \{(g_i, b_j) \mid g_i \in G, b_j \in B\}$, where $B = \{\text{present, absent}\}$.

```
procedure GENEDETECTION(R, k, θ, 𝒢)
    for each r ∈ R do                                              ▷ Read processing
        mid_kmer ← GETMIDKMER(k, r)                                ▷ (i) Filtering
        if mid_kmer ∉ 𝒜 then
            continue
        freqs, hits ← []                                           ▷ (ii) k-mer counting
        K ← GETALLKMERS(k, r)
        for each kmer ∈ K do
            (matched_gene, hit_pos) ← FIND(kmer, 𝒢)
            if matched_gene ≠ ∅ then
                freqs[matched_gene] ← freqs[matched_gene] + 1
                hits[matched_gene].add(hit_pos)
    prob_seqs ← []                                                 ▷ (iii) Selection of probable sequences
    for each gene ∈ G do
        if freqs[gene] > 0 then
            prob_seqs.add(gene)
    (depths, coverages) ← INITDEPTHSANDCOVERAGES(G, prob_seqs)
    for each gene ∈ prob_seqs do                                   ▷ (iv) Depth and coverage calculation
        depths[gene] ← CALCULATEDEPTH(hits[gene])
        coverages[gene] ← CALCULATECOVERAGE(depths[gene])
    for each allele ∈ prob_seqs do                                 ▷ (v) Reporting
        if coverages[gene] ≥ θ then
            print gene + "present"
        else
            print gene + "absent"
```

Figure 7. Detailed STing gene detection algorithm. Input of this algorithm comprises the sequencing reads to process, *k*-mer size, minimum coverage threshold to consider a gene as present in the sample, a list of the genes of interest to be searched in the sample, and the gene ESA index. The algorithm output corresponds to the list of genes with *k*-mer matches and their corresponding tag for presence or absence.

2.3.5 *Genomic data for sequence typing.*

We used 1,050 Illumina sequencing read sets of isolates from four bacterial species (Campylobacter jejuni, Chlamydia trachomatis, Neisseria meningitidis, and Streptococcus pneumoniae) retrieved from the PubMLST[5]/EBI ENA[6] database to execute the experiments (Table 14). Using the isolate metadata available on PubMLST, we selected 40 samples from the four species (10 samples each) for the MLST comparative test, and 20 samples of N. meningitidis isolates for the larger typing schemes (rMLST, and cgMLST) comparative test. We selected these two datasets trying to capture the diversity of the most common STs of each species in the PubMLST database and preferring recently sequenced isolates. For the large-scale accuracy test, we used a dataset of 1,000 samples of N. meningitidis isolates.

2.3.6 *Computational environment.*

We used a machine provided with RedHat Linux SO, 24 cores, and 64 GB of RAM to perform the experiments described in this study.

2.3.7 *MLST comparative test design.*

To measure the performance of our application on the traditional seven loci MLST analysis, we compared STing (v0.24.2) in two execution modes, fast and sensitive, along

[5] https://pubmlst.org/
[6] https://www.ebi.ac.uk/ena

with six applications able to perform sequence typing (stringMLST (Gupta, et al., 2017), MentaLiST (Feijao, et al., 2018), Kestrel (Audano, et al., 2018), SRST2 (Inouye, et al., 2014), ARIBA (Hunt, et al., 2017), and Offline CGE/DTU). These applications can be classified into five groups depending on the strategy (algorithmic paradigm) used to predict the sequence types of whole genome sequencing data samples from bacterial isolates: *k*-mer, *k*-mer plus alignment, mapping, mapping plus local assembly, and assembly (Table 4). For the Offline CGE/DTU application, we used the script `runMLST.py`[7] (Pritchard, 2014) , an offline implementation of the original alignment-based MLST method from the Center of Genomic Epidemiology[8] (CGE) (Larsen, et al., 2012). This implementation uses multithread BLAST searching for the MLST analysis, as opposed to STing, which is a single thread application. To fairly compare STing with the Offline CGE/DTU implementation, we modified the script `runMLST.py` to use only one thread for BLAST searches. For each application, we measured the accuracy in terms of percentage of alleles correctly predicted from the total samples analyzed, and the performance in terms of average run time and average peak of RAM required to analyze each of the 40 samples in the dataset. We reported the average run time and average max RAM as the average of three executions of each application per sample analyzed. Kestrel requires the generation of a *k*-mer counts file before it can be run to predict STs. For this purpose, we used the application KAnalyze (Audano and Vannberg, 2014) (v2.0.0) with the parameters as described (Audano, et al., 2018). We reported the average run time of Kestrel as the sum of the average times of KAnalyze and Kestrel for processing each sample, and the average RAM consumption as the maximum average peak of RAM consumed by the two

[7] https://github.com/widdowquinn/scripts/blob/master/bioinformatics/run_MLST.py
[8] http://www.genomicepidemiology.org/

applications on each sample. Since the Offline CGE/DTU application requires complete assemblies to predict STs, we assembled each isolate read sample using the application SPAdes (Bankevich, et al., 2012) (v3.13.0) with default parameters. We reported the average runtime as the sum of the average times of SPAdes and Offline CGE/DTU to process each sample, and the average RAM consumption as the maximum average peak of RAM consumed between the two applications during the analysis of each sample. The commands used with each application tested are listed in Table 5.

Table 4. Sequence typing applications tested.

Application	Algorithm Type[a]	Input Type	Version	Reference[b]
STing	*k*-mer	Reads	0.24.2	DOI 10.1101/855478
stringMLST	*k*-mer	Reads	0.6.1	PMID 27605103
MentaLiST	*k*-mer	Reads	1.0.0	PMID 29319471
Kestrel	*k*-mer + alignment	Reads	1.0.2dev1	PMID 29186321
SRST2	Mapping	Reads	0.2.0	PMID 25422674
ARIBA	Mapping + assembly	Reads	2.13.3	PMID 29177089
Offline CGE	Assembly	Assembly	-	https://github.com/widdowquinn/scripts/blob/master/bioinformatics/run_MLST.py; Implementing method from PMID 22238442

[a] Algorithmic paradigm implemented by the tool
[b] Reference of the software application

Table 5. Commands used with each sequence typing software.

Application	Task	Command
STing	DB creation	indexer -c <config_file> -p <db_prefix>
	Sequence typing (fast)	typer -x <db_prefix> -1 <fastq_1> -2 <fastq_2> -k 30 -c -s <sample_name>
	Sequence typing (sensitive)	typer -x <db_prefix> -1 <fastq_1> -2 <fastq_2> --sensitive -s <sample_name> -k 30 -n 3 -c -a -d -t <out_depth_file>
stringMLST	DB creation	stringMLST.py --getMLST --species <species_name> -k 35 -P <db_prefix>
	Sequence typing	stringMLST.py --predict -1 <fastq_1> -2 <fastq_2> -k 35 -P <db_prefix>
MentaLiST	DB creation	mentalist build_db -k 30 --db <out_db_file> -p <profile.txt> -f <fasta_files>
	Sequence typing	mentalist call --db <db_file> -o <out_file> -1 <fastq_1> -2 <fastq_2>
Kestrel	*k*-mer counts file creation (KAnalyze)	java -Xmx3G -jar kanalyze.jar count -k 31 --countfilter=kmercount:5 --quality=10 -m ikc --minsize 15 -o <kanalyze_out_File> <fastq_1> <fastq_2>
	Sequence typing	java -Xmx2G -Dlogback.configurationFile=logback.xml -jar kestrelmlst.jar <allele_seqs_file> <kanalyze_out_file> > <kestrel_mlst_out_file>
SRST2	DB preparation (Samtools and Bowtie2)	samtools faidx <allele_seqs_file>; bowtie2-build <allele_seqs_file> <allele_seqs_file>
	Sequence typing	srst2 --output <out> --input_pe <fastq_1> <fastq_2> --mlst_db <allele_seqs_file> --mlst_definitions <profile_file> --mlst_delimiter '_'
ARIBA	DB creation	ariba pubmlstget <species_name> <out_dir>
	Sequence typing	ariba run <db_ref_dir> <fastq_1> <fastq_2> <out_dir>
Offline CGE	Assembly (SPAdes)	spades.py -1 <fastq_1> -2 <fastq_2> -o <out_dir> -t <threads>
	Sequence typing	run_MLST.single_thread.py -o <out_dir> -i <scheme_dir> -g <genomes_dir> -p <profile> -l <log_file> -v --force

2.3.8 Large-scale MLST accuracy test design.

To measure the accuracy of our application using the MLST scheme on a large-scale dataset, we ran STing in fast mode on 1,000 samples of *N. meningitidis*. We measured the accuracy in terms of the percentage of STs correctly predicted from the total samples analyzed, and the performance in terms of average run time and average peak of RAM required to analyze each of the 1,000 samples of the dataset. We reported the average run time and average max RAM as the average of five executions of the application per sample analyzed.

2.3.9 Limit of detection, and performance on single and multithread environment test design.

We evaluated the minimum sequencing depth required for correctly predicting STs on whole genome sequencing samples from bacterial isolates. We retrieved 1,306 assemblies of *Campylobacter jejuni* (n=581) and *Neisseria meningitidis* (n=725) with known MLST information from the GenBank[9] database (Table 15). Then, we simulated Illumina paired-end reads – HiSeq 2500, 2x150 bp, 500bp of average fragment length, with 10 as the fragment size standard deviation – from each genome at seven sequencing depths (1, 3, 5, 10, 15, 20, and 40x) using the software ART (Huang, et al., 2012) (v2.5.8). We executed STing (fast mode) over each generated sample to measure the accuracy in terms of the percentage of correct STs and alleles predicted from the total samples at each sequencing depth. We also evaluated the performance of STing in multithread environments. We executed 20 parallel instances of STing to analyze the 1,306 samples and measure the average time required to process the complete dataset at each sequencing depth.

[9] https://www.ncbi.nlm.nih.gov/genbank/

2.3.10 Large-scale sequence type schemes comparison test design.

To evaluate the scalability, accuracy, and performance of our application on large-scale sequence typing schemes, we compared STing (fast and sensitive modes) on 20 samples of *N. meningitidis* against other sequence typing applications using the rMLST (loci=53), and the cgMLST (loci=1,605) schemes. We used three applications (stringMLST, SRST2, and Offline CGE) for rMLST, and three applications (stringMLST, MentaLiST, and Offline CGE) for cgMLST, which were able to execute the sequence typing analysis successfully using these larger schemes. For each application and typing scheme, we measured the accuracy in terms of the percentage of correct allele predictions from the total alleles of the tested samples, and the performance in terms of the average of run time and max peak of RAM required to process each sample from the dataset.

2.3.11 Gene detection test design.

We evaluated the ability of STing to predict the presence/absence of sequences of interest in NGS read samples by detecting antimicrobial resistance (AMR) genes and virulence factor (VF) genes in simulated Illumina read datasets. We retrieved 71 assemblies from the GenBank database that correspond to 25 species listed in the World Health Organization priority list of antibiotic-resistant bacteria and tuberculosis (Tacconelli, et al., 2018) (Table 16). Then, we simulated Illumina paired-end reads – HiSeq 2500, 2x150bp, 500bp of average fragment size, with 10 as the fragment size standard deviation – from each genome at 20x and 40x sequencing depth, using the software ART. For the AMR gene detection test, we used 1,434 AMR genes available in the Comprehensive Antibiotic Resistance Database (CARD, v2.0.2) (Jia, et al., 2017). For the VF gene detection test, we used 1,443 genes from the virulence factor database (VFDB, release date 03-22-2019) (Liu, et al., 2019). In both tests, we first defined presence/absence of each gene in each genome using BLASTn (v2.2.28+) (Camacho, et al., 2009), as a ground-truth for assessing the STing performance. To perform a fair comparison with STing's gene detection, which is based is based on exact pattern matching, we defined a cutoff of 100% for identity and query (gene) coverage in BLASTn to consider a gene as present in a genome, *i.e.*, if the gene is perfectly contained in the genome. Then, we built databases on STing for each gene set of interest (CARD and VFDB), and executed the respective gene detection analysis on each genome-derived read set at each sequencing depth, using a threshold of 100% for gene coverage to consider a gene as present in a sample. Finally, we evaluated the performance of detection in terms of sensitivity, specificity, precision, and accuracy, which are defined as follows:

$$Sensitivity = \frac{TP}{TP+FN};$$

$$Specificity = \frac{TN}{TN+FP};$$

$$Precision = \frac{TP}{TP+FP};$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN};$$

where, $TP =$ true positives, $TN =$ true negatives, $FP =$ false positives, and $FN =$ false negatives.

2.4 Results and discussion

We compared STing to six of the most widely used programs for genome-enabled molecular typing, including its predecessor stringMLST (Figure 8). The programs were evaluated for accuracy in terms of the percentage of correct allele predictions, speed in terms of average run time, and efficiency in terms of average maximum RAM consumption. Genome-enabled typing programs can be classified according to the algorithmic paradigm that they use: *k*-mer only, *k*-mer plus alignment, read-to-genome mapping, mapping with local assembly, and full assembly. STing uses the minimalist *k*-mer only approach. STing was run in the fast and sensitive modes for the traditional housekeeping MLST scheme and two larger-scale typing schemes, rMLST and cgMLST. Allele databases for all three typing schemes were taken from the PubMLST database (https://pubmlst.org/). The STing fast mode uses a *k*-mer matching only strategy, and the sensitive mode includes an additional step whereby false positive matches are excluded based on gaps in the coverage profiles of *k*-mer matches to allele sequences. Comparisons were performed for 10 samples each across four species that are widely used in MLST and accordingly have diverse MLST databases: *Campylobacter jejuni*, *Chlamydia trachomatis*, *Neisseria meningitidis*, and *Streptococcus pneumoniae*. STing shows 100% accuracy, in both the fast and sensitive modes, as well as the fastest run time and lowest memory use of any program for MLST (Figure 8A). The results of the same comparisons are broken down for each of the four individual species in Figure 9. We also ran STing for MLST across a range of sequence coverage levels in an effort to assess its detection limits and multi-core performance (Figure 10A). STing performs best at 40x coverage, but it maintains accuracy at 20x with a marginal drop-off at 10x. While STing is designed as a single core application, we found that executing multiple threads of the program allows it to maintain run time up to 40x coverage (Figure 10B). This provides for a straightforward way to run STing on numerous genome samples; the MLST accuracy and speed metrics for STing run on a larger dataset of 1,000 *N. meningitidis* samples are shown in Table 6. When this large-scale analysis was performed, STing was able to uncover seven samples that were initially scored as erroneous predictions but actually turned out to be mis-annotated on the PubMLST database (Table 7). STing also shows the highest accuracy, speed, and efficiency, for the four programs that are capable of genome-enabled rMLST typing (Figure 8B). Programs that show an 'X' in these comparisons were unable to run for a variety of reasons related to their initial design, the runtime, and database indexing limitations. The program Mentalist shows marginally higher accuracy, run time, and efficiency for cgMLST compared to STing, which shows the second best metrics for these categories (Figure 8C). However, the utility of Mentalist, which was designed specifically for cgMLST, is limited by the size of the database that can be indexed. For that reason, it could not be run on the latest rMLST database available from PubMLST.

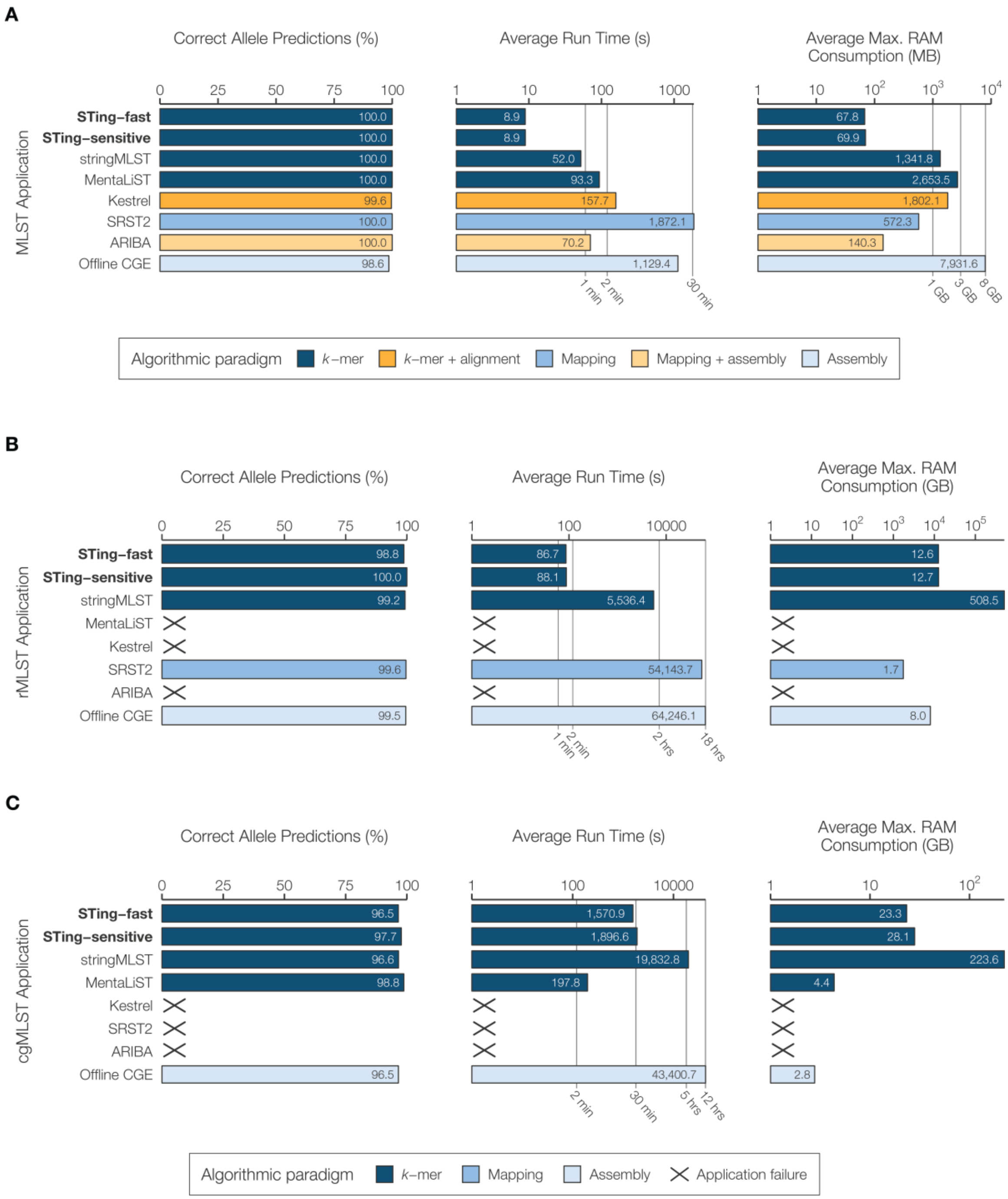


Figure 8. Performance comparison of STing with six other sequence typing applications. The fast and sensitive modes of STing are compared to 6 other contemporary typing applications to measure the accuracy and runtime performance, using three different typing schemes: (A) the traditional MLST (loci=7) on 40 samples from four bacterial species (10 samples per species: *C. jejuni*, *C. trachomatis*, *N. meningitidis*, and *S. pneumoniae*); (B) the ribosomal MLST (rMLST) scheme (loci=53) on 20 samples of *N. meningitidis*, and (C) the core genome MLST (cgMLST) scheme (loci=1,605) on 20

samples of *N. meningitidis*. The typing applications are color-coded based on the algorithmic paradigms that they utilize for performing sequence typing. Performance is measured in terms of the percentage of correct alleles predicted, the average runtime across each dataset measured in seconds (displayed in log-scale), and average peak RAM utilization across each dataset measured in megabytes (MB) for MLST, and gigabytes (GB) for rMLST and cgMLST (both displayed in log-scale).

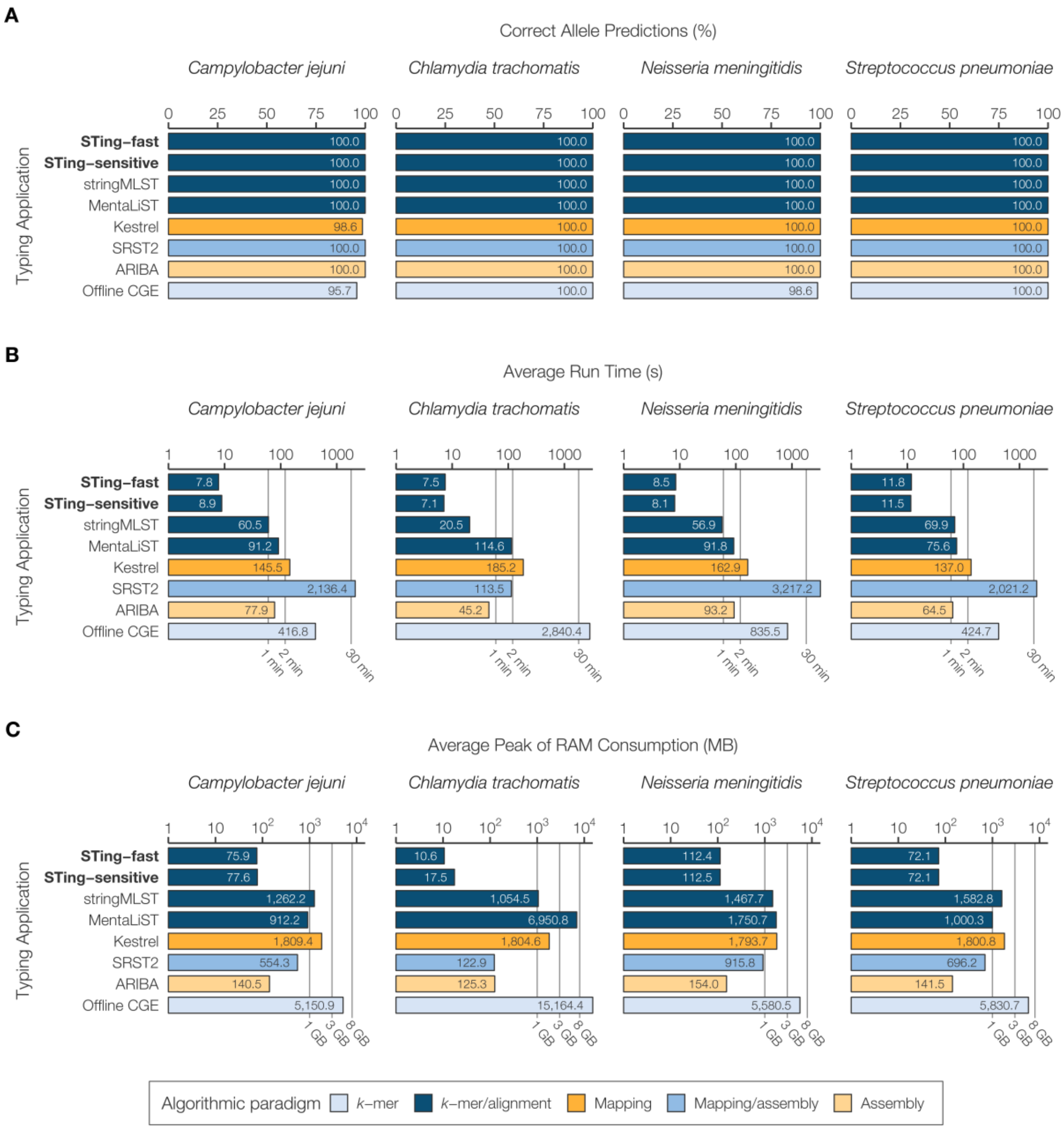


Figure 9. Performance comparison of STing for MLST detailed by species. The fast and sensitive mode of STing is compared to six other contemporary typing applications to measure the accuracy and runtime performance using the traditional MLST (loci=7) on 40 samples from four bacterial species (10 samples per species): *C. jejuni*, *C. trachomatis*, *N.*

meningitidis, and *S. pneumoniae*. Performance is measured in terms of (A) the percentage of correct alleles predicted, (B) the average runtime across each dataset measured in seconds (displayed in log-scale), and (C) the average peak RAM utilization across each dataset measured in megabytes (MB) for MLST, and gigabytes (GB) for rMLST and cgMLST (both displayed in log-scale). The typing applications are color-coded based on the algorithmic paradigms that they utilize for sequence typing

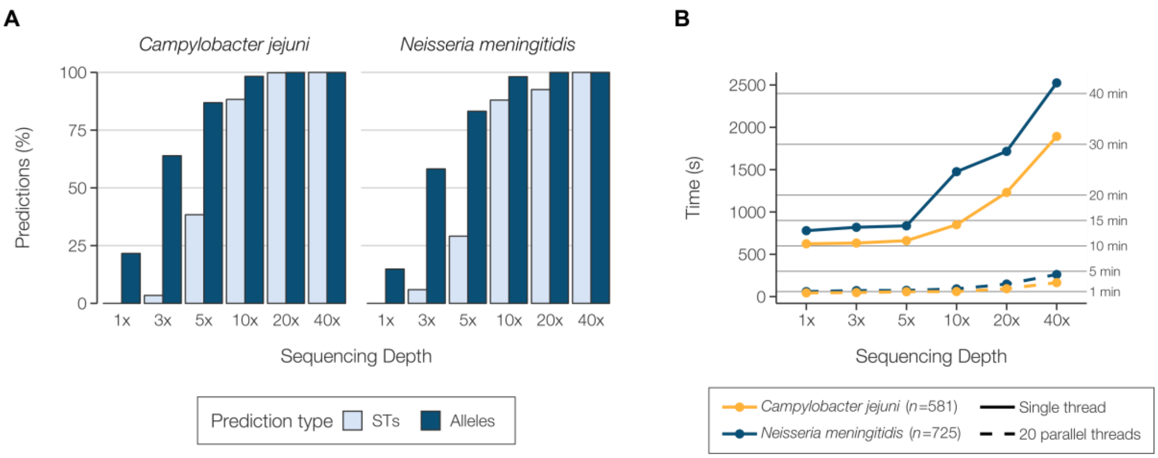


Figure 10. Results of the limit of detection test, and single- and multi-core performance test. STing's typer utility was run in the fast mode over 1,306 read samples simulated at six sequencing depths (1, 3, 5, 10, 20, an 40x) from assemblies of two species *Campylobacter jejuni* (n=581) and *Neisseria meningitidis* (n=725). (A) Percentage of correct predictions in terms of STs and alleles at different sequencing depths for the two datasets. (B) Total time in seconds required to process the complete dataset for each species at different sequencing depths using a single thread or instance (solid lines) and 20 multiple threads of the typer utility.

Table 6. Results of the MLST benchmarking test on a large-scale dataset.

Species	#Sam.[b]	Depth[d]	U. STs[c]	#All.[a]	Mode[e]	Correct Alleles[f] (%)	Time[g] (s)	RAM[h] (MB)	Proc. Rate[i] (MB/s)	Proc. Rate[j] (Reads/s)
N. meningitidis	1,000	175.1	387	7,197	Fast	100.0	10.6	82.9	86.2	372,739
					Sensitive	100.0	11.4	82.9	80.7	348,760

[a] Total number of alleles tested
[b] Total number of samples tested
[c] Total number of unique STs in the dataset
[d] Average sequencing read depth of the dataset
[e] Execution mode of the typing tool
[f] Percentage of correctly predicted alleles
[g] Average run time in seconds per sample
[h] Average of maximum RAM usage per sample
[i] Data processing rate in megabytes per second
[j] Data processing rate in reads per second

Table 7. List of samples correctly predicted by STing and misannotated in PubMLST.

Accession[a]	Predicted ST	ST on PubMLST
ERR036115	11	672
ERR133727	106	1087
ERR133738	116	106
ERR133744	10307	989
ERR137178	144	102
ERR310540	11	5757
ERR957622	154	6697

[a] Accession number of the isolate in the EBI ENA database

In addition to molecular sequence typing, STing can also be used for automated gene detection directly from NGS reads. The gene detection mode uses a database of genes of interest, and we used databases of AMR and VF genes given their public health relevance. The Comprehensive Antibiotic Resistance Database[10] (CARD) of 1,434 AMR genes and the Virulence Factors of Pathogenic Bacteria database[11] (VFDB) of 1,443 VF genes were used for this purpose (Jia, et al., 2017; Liu, et al., 2019). STing was used to query the AMR and VF databases with 71 NGS genome datasets for 25 bacterial pathogen species taken from the World Health Organization global priority list of antibiotic-resistant bacteria (Tacconelli, et al., 2018). STing shows very high accuracy metrics for both AMR and VF detection (Figure 11A), along with fast and efficient performance (Figure 11B). STing can be run in in this way to rapidly detect any genes of interest, which extends its utility beyond public health genomics. This could be particularly useful for large scale environmental genomics.

[10] https://card.mcmaster.ca/
[11] http://www.mgc.ac.cn/VFs/

A

		Antimicrobial Resistance				Virulence Factor			
Seq. Depth:		20x		40x		20x		40x	
		STing				STing			
		Present	Absent	Present	Absent	Present	Absent	Present	Absent
Known	Present	132 TP	0 FN	132 TP	0 FN	1,166 TP	4 FN	1,170 TP	0 FN
	Absent	3 FP	101,679 TN	4 FP	101,678 TN	10 FP	101,273 TN	10 FP	101,273 TN

	AMR 20x	AMR 40x	VF 20x	VF 40x
Sensitivity:	1.000	1.000	0.996	1.000
Specificity:	1.000	1.000	1.000	1.000
Precision:	0.978	0.971	0.991	0.992
Accuracy:	1.000	1.000	1.000	1.000

B

Test[a]	Seq. Depth	Avg. Time[b] (s)	Avg. RAM[c] (MB)	Total Time[d] (s)
Antimicrobial Resistance	20x	3.7	71.8	200.1
	40x	8.0	72.4	400.4
Virulence Factor	20x	3.9	79.0	460.3
	40x	12.7	134.0	900.4

[a] Type of gene detection test executed
[b] Average time in seconds required to process each genome read sample
[c] Average maximum RAM required to process each genome read sample
[d] Total time in seconds required to process the entire read dataset

C

Feature	STing	stringMLST	MentaLiST	Kestrel	SRST2	ARIBA	Offline CGE
Assembly- & alignment-free	✓	✓	✓				
Standalone	✓	✓	✓				
MLST support	✓	✓	✓	✓	✓	✓	✓
Larger schemes support	✓	✓	✓		✓		✓
Runtime ≤ 1 hour for larger schemes	✓	✓	✓				
Big DBs support (≥ 2^{15} alleles per locus)	✓	✓					✓
k-mer size independent DB	✓				–	–	–
Gene detection	✓				✓	✓	
Confidence information	✓		✓				
Automated DB download	✓	✓	✓		✓	✓	

✓ Feature included
– Not applicable

Algorithmic paradigm: *k*–mer (STing, stringMLST, MentaLiST); *k*–mer + alignment (Kestrel); Mapping (SRST2); Mapping + assembly (ARIBA); Assembly (Offline CGE)

Figure 11. Performance comparison of STing's Gene Detection program. STing's Gene Detection program was run on 71 WHO designated high-priority bacterial genomes (simulated at a read depth of 20x and 40x) that contained gene annotations for 1,434 AMR genes and 1,443 VF genes. Confusion matrices for the detection of (A) AMR genes from the CARD dataset, and VF genes from the VFDB dataset are shown. (B) The table demonstrates the accuracy and average runtime performance comparison of STing's Gene Detection at each sequencing read depth. (C) Feature comparison between STing and the six applications tested for sequence typing.

CHAPTER 3. WEBSTING: A RAPID AND ACCURATE ALIGNMENT-FREE WEB PLATFORM FOR BACTERIAL PATHOGEN CHARACTERIZATION

3.1 Introduction

Next generation sequencing (NGS) technologies are now a primary tool in molecular epidemiology due to their wide availability, low cost, throughput, and speed. Public health agencies around the world are increasingly relying on NGS technologies for surveillance and control of infectious diseases. As a result, public health laboratories are producing an enormous amount of sequence data, and challenges have emerged for the analysis and administration of such massive amounts of information (Fricke and Rasko, 2014). Both analysis and administration of NGS data on a large scale require substantial bioinformatics-related infrastructure and expertise. These resources are expensive and are usually only available to well-funded public health institutions.

Cloud-based platforms are a suitable alternative to institutions that lack the computational capacity required for the bioinformatics analysis of NGS data produced for routine molecular epidemiology. Solutions like Cloud BioLinux and CloVR facilitate bioinformatic analysis of an enormous quantity of biological data with virtual machines on cloud service providers (Angiuoli, et al., 2011; Krampis, et al., 2012). Even when the cloud approach that offers both software and infrastructure on-demand is very convenient, deploying the required applications for custom bioinformatics workflows require expertise and considerable time and effort.

As a solution to the problem of software deployment, Web-based bioinformatics platforms offer a wide range of programs as ready-to-use services. Platforms like Galaxy (Afgan, et al., 2018) provide various bioinformatics programs, including sequence alignment, read mapping and assembly, gene prediction, and genome annotation. Under this model, users do not need to deploy software but must select and connect the proper services to build specific automated workflows for data analysis, which also requires bioinformatics expertise. For example, designing a classic alignment-based automated pipeline on these platforms for characterizing bacterial pathogens from WGS data requires selecting the appropriate application among several available for each analysis step, *e.g.*, read quality control, *de novo* read assembly, sequence typing, and gene annotation. In the context of NGS for molecular epidemiology, a pressing need exists for specialized and automated platforms for bacterial pathogen analysis.

Bioinformatics platforms based on NGS technologies have been developed specifically for public health microbiology. For instance, Nullarbor[12] is a platform developed by the Microbiological Diagnostics Unit Public Health Laboratory from the University of Melbourne for the automated analysis of sequenced bacterial isolates that includes quality control, species identification, genome assembly, and characterization by sequence typing and gene profiling. However, this platform is a command-line software solution that requires mid to high levels of bioinformatics skills for operation. Another example is Pathogenwatch[13] (formerly known as WGSA), a Web-based platform for genomic surveillance developed by the Centre for Genomic Pathogen Surveillance[14]

[12] https://github.com/tseemann/nullarbor
[13] https://pathogen.watch/
[14] https://www.pathogensurveillance.net/

(CGPS), offers an extensive database of bacterial genomes and allows the automated characterization of user-submitted assemblies and WGS raw samples. The last example is the platform from the CGE, which offers several analyses including sequence typing, phenotyping, phylogeny, and an automated pipeline for pathogen characterization (Thomsen, et al., 2016). Although the last two described platforms are promising and offer easy Web access, they have limitations. In the case of Pathogenwatch, all the services offered are based on sequence alignment and require an already assembled genome or the raw WGS reads to perform *de novo* assembly. This method is not well suited for real-time molecular epidemiology on a large scale because it requires considerable computational resources and time required for *de novo* assembly. Regarding the platform from the CGE, even when some services offered are based on alignment-free algorithms, the pipeline for automated pathogen characterization requires genome assembly and is not currently maintained.

New dedicated Web-based solutions for genome-enabled molecular epidemiology must facilitate access to the public health community to methods for rapid WGS data analysis. This chapter presents WebSTing, a Web-based platform for the automated characterization of bacterial pathogens that uses the innovative alignment-free algorithm STing for rapid and accurate analysis of WGS samples.

3.2 Materials and methods

3.2.1 Architecture

The architecture of the current implementation of WebSTing is shown in Figure 12. WebSTing supports three types of clients: web, mobile devices, and batch mode. WebSTing implements a frontend Application Program Interface (API) for data retrieval and displays for the three types of clients. WebSTing comprises a two main service tiers: (1) Web and (2) STing. The Web services tier interacts directly with users through a load balancer, a Web server, a user interface, and the frontend API. It is also responsible for managing the transfer and storage of read sample files. The STing services tier implements the typing and gene detection features of STing as services that consume the read files for processing and communicate results to the Web services tier.

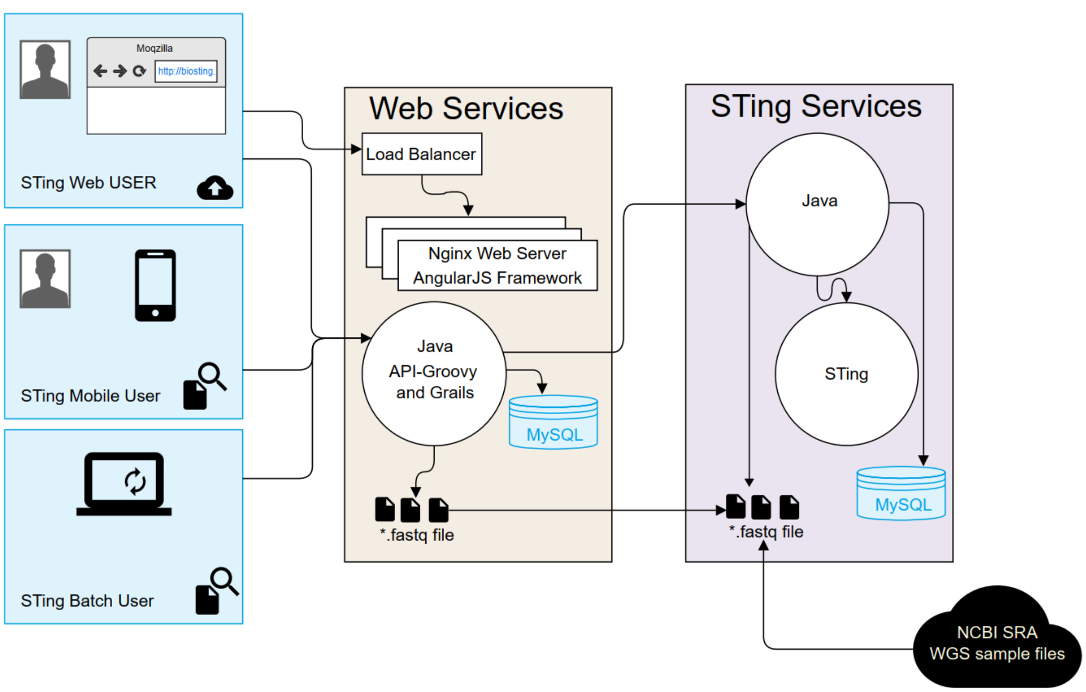


Figure 12. The architecture of the current implementation of WebSTing.

3.2.2 Development technologies

The Web server is implemented with Nginx, and the user interface uses the JavaScript frameworks Angular, Bootstrap, and JQuery. Authentication and access control are based on the Spring Security framework. The frontend API is implemented using the Web development framework Grails with elements written in Java and Groovy programming languages. The database logic relies on domain classes implemented with the Hibernate framework and uses the MySQL database management system.

3.2.3 Frontend REST API

Services provided by the frontend REST API are used by users to display data to any client. Table 8 lists the API methods for retrieving data from WebSTing.

3.2.4 Scalable cloud-based design

Scalability is a crucial feature required by a Web-based platform dedicated to supplying services for real-time molecular epidemiology. WebSTing considers a cloud-centric design that allows for easy and rapid scalability at the level of computation power, memory, and storage. This design also allows for a smooth deployment on major cloud service providers such as Amazon Web Services, Google Cloud, or Microsoft Azure. The cloud-based design of WebSTing uses Docker containers for distributing the different system tiers: Web server and user interface, frontend API, and the STing services. Figure 13 shows the cloud-based design of WebSTing.

Table 8. API method definition.

API Method	Description	End Point	Request	External Server Call
processOrganismTSV	Load all organisms into database from static file organism-type-schemes.tsv	<servername>:8080/uploadData	GET	
compareGetAll	Returns 1 completed organism record along with a list of alleles that STing produced during processing	<servername>:8080/compareGetAll	GET	/STingWebService/service/stingResult/compare/<organismName>/<schemaType>
retrieveAccessionFileGetAll	Returns all records added to the system for lookup and includes lookup status		GET	/STingWebService/service/retrieveAccessionFiles/getAll
retrieveAndInsertAccessionFile	Inserts 1 record into the retrieved table to be queued for retrieval		GET	/STingWebService/service/retrieveAccessionFiles/insert/<accessionName>/<schema>/<organismName>
getAllOrganismType	Get all organism type records from the database	<servername>:8080/getAllOrganismType	GET	
stingUpload	Inserts record into upload table to be queued for file copy from web server	<servername>:8080/stingUpload	GET	/STingWebService/service/stingUpdate/upload/<filename1>/<filename2>/<stingId>
getCompareSample	From the ID list, the samples are compared, and the result is a tree PNG file	<servername>:8080/getCompareSample	GET	/STingWebService/service/stingResult/compareSamples/<samples>
getAll	Returns all records that have been processed and their status	<servername>:8080/getAll	GET	/STingWebService/service/stingResult/getAll
uploadData	Using a chunking method that uploads large file	<servername>:8080/uploadData	POST	

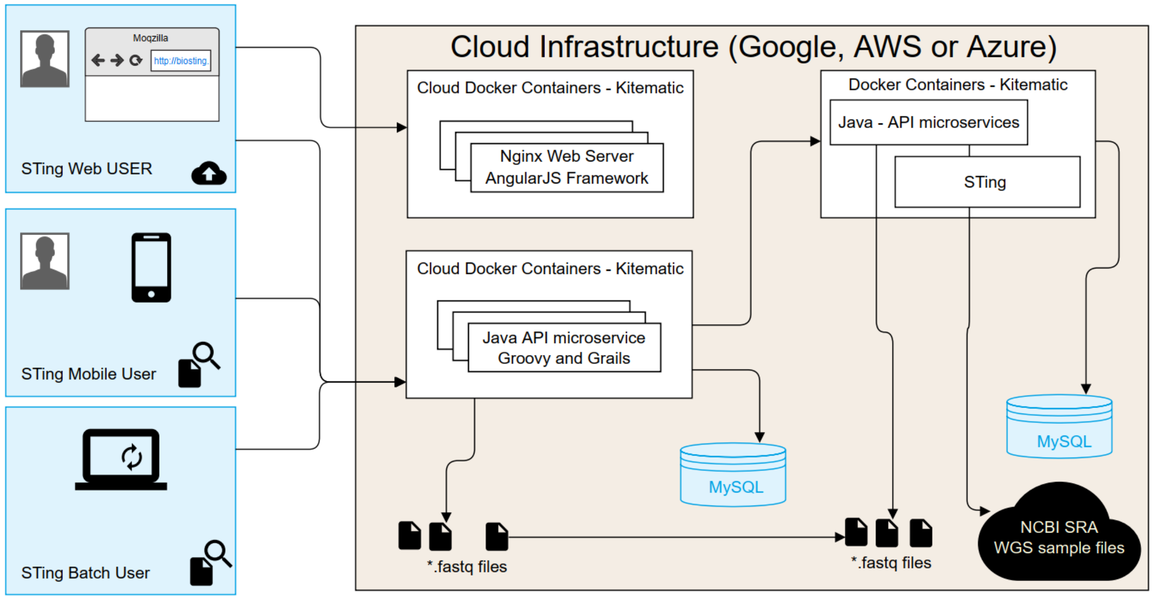


Figure 13. WebSTing architecture design for full scalability on Cloud infrastructure.

3.3 Results and discussion

WebSTing is the result of a joint effort between academia and industry, with the Jordan Lab from the School of Biological Sciences of the Georgia Institute of Technology, and the Applied Bioinformatics Laboratory (ABiL), to create a turn-key solution for molecular epidemiology in the form of a modern Web-based platform with the highest standards of software development. Such collaboration, allowed us to combine one of the state-of-the-art alignment-free methods for NGS-based molecular epidemiology analysis – STing, along with the newest development technologies to produce a reliable, secure, and scalable Web platform for the automated characterization of bacterial pathogens using WGS data.

WebSTing is an easy-to-use Web platform for the rapid characterization of bacterial pathogens that includes sequence typing, antimicrobial resistance gene detection, and phylogenetic analysis. WebSTing is a multi-user platform that supplies industry-standard authentication and access control and full cloud-based scalability. WebSTing can be deployed on any of the major cloud service providers like Amazon Web Services, Google Cloud, or Microsoft Azure and can be easily configured to scale the requirements in terms of number of users, computing power, memory, and storage space.

WebSTing's services are available through user registration. Once users register, they can run any of the tasks for pathogen characterization from the root page of the platform –Dashboard (Figure 14). Users can upload WGS read samples (FASTQ files) or supply accession numbers of samples to be retrieved automatically from the Sequence Read Archive (SRA) database of the National Center for Biotechnology Information (NCBI). When WGS read samples are loaded to the platform, users can perform sequence typing, using either the classic Multilocus Sequence Typing (MLST) or the core genome MLST (cgMLST) schemes, and antimicrobial resistance (AMR) gene detection using one of four available reference databases. WebSTing preserves a history of results from all the analysis executed for each user. In the case of sequence typing, this history allows for comparing multiple samples previously analyzed to find evolutionary relationships via phylogenetic analysis. The following section describes each of the analyses available on WebSTing.

A
B

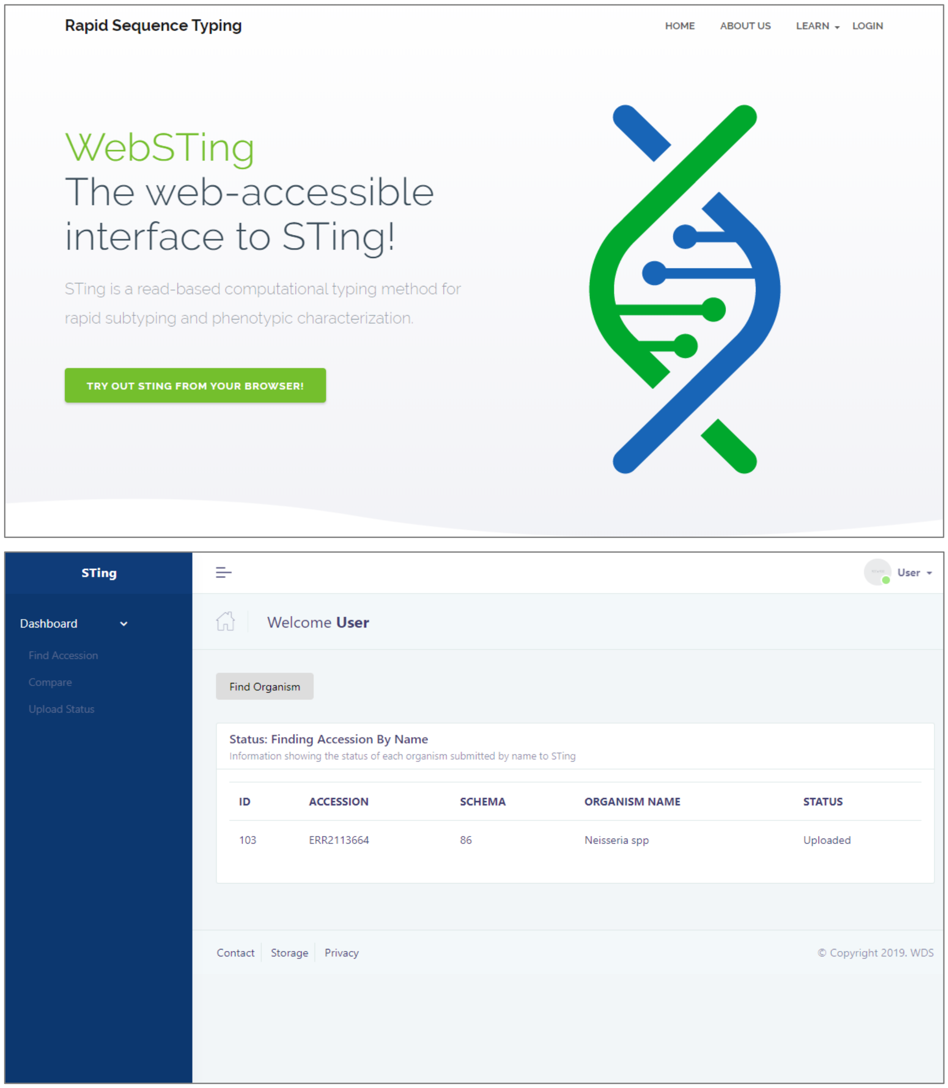


Figure 14. WebSTing platform home page and Dashboard. (**A**) WebSTing home page summarizes the platform features and allows users to register or login to the platform. (**B**) WebSTing Dashboard allows users to start new analyses of sequence typing, gene detection, or compare samples through phylogenetic trees.

3.3.1 Sequence typing

The sequence typing analysis uses the STing's typer utility as a service. The user first selects the scheme or type of analysis to be run (Figure 15). Available scheme options for sequence typing are the classic MLST (seven to ten loci) and the core genome MLST (cgMLST, ~1,000> loci). Then, the user selects the organism-specific database. Databases for both MLST and cgMLST are periodically updated. Weekly, WebSTing retrieves the latest version of all available databases from the PubMLST website (http://pubmlst.org/) and transforms them into ESA indexes using the STing indexer utility. Currently, 147 databases are available for the MLST (n = 141) and cgMLST (n = 6) schemes on WebSTing. After selecting the database, the user provides the WGS sample for analysis, through entering the corresponding accession number from the SRA database (Figure 15A), or by uploading the sample read files (Figure 15B). Then, the user is notified about the file transfer status for each sample on the Dashboard (Figure 16A). When sample files are transferred to WebSTing, the sequence typing process starts automatically. Once the sequence typing has finished, results are available on the "Processing Status" page (Figure 16B). Sequence typing results include the predicted ST for each sample, the total *k*-mers and reads processed, and a list of alleles called for each locus of the organism-specific typing scheme with their corresponding sequence length and coverage percentage (Figure 17).

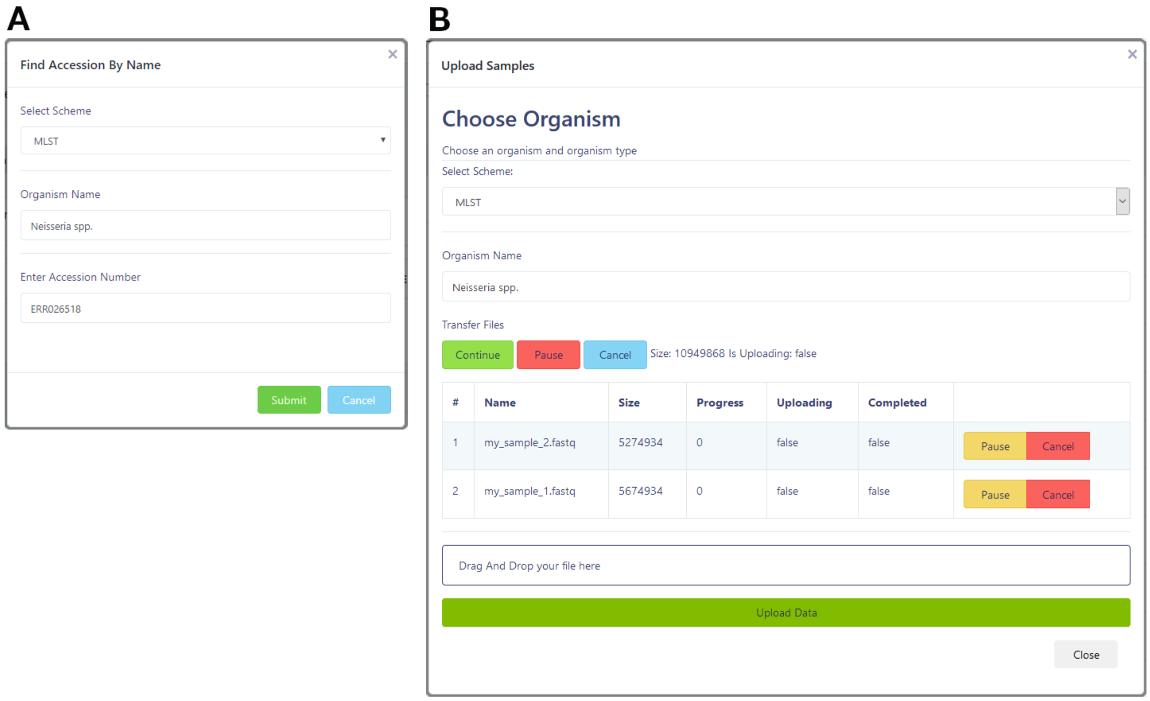


Figure 15. Selection of analysis scheme, species, and samples to analyze. To start characterizing a pathogen it is required to select the scheme of type of analysis (MLST, cgMLST, or antimicrobial resistance –AMR– gene detection), the database of reference (specific for an organism in the case of sequence typing), and the read sample. The read sample can be specified by (**A**) entering an accession number from the SRA database, or (**B**) uploading the read files.

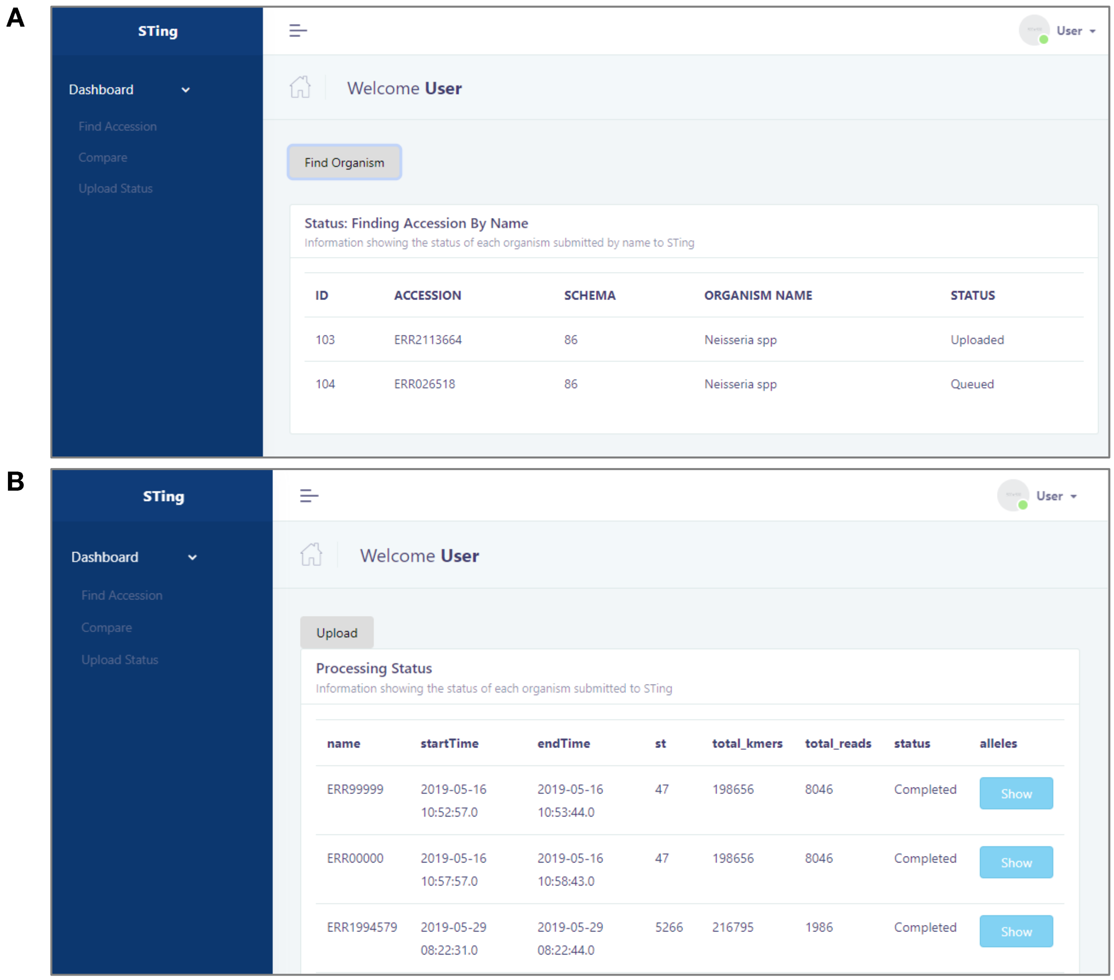
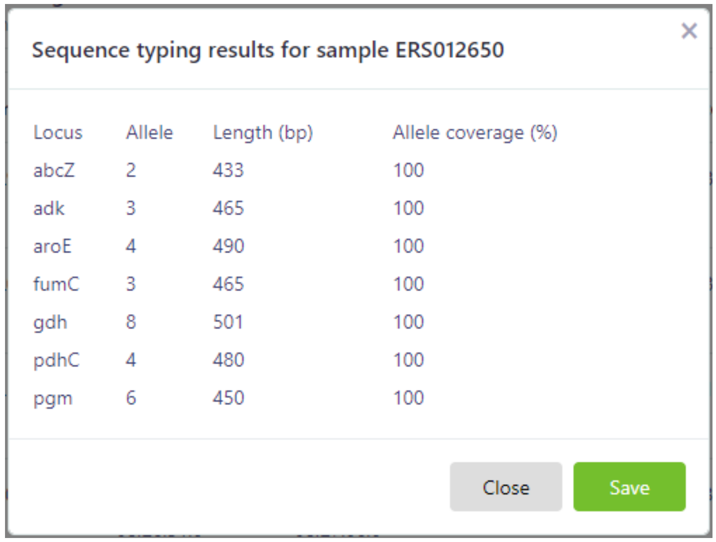


Figure 16. Interfaces for checking the transfer and processing status of samples.


Sequence typing results for sample ERS012650

Locus	Allele	Length (bp)	Allele coverage (%)
abcZ	2	433	100
adk	3	465	100
aroE	4	490	100
fumC	3	465	100
gdh	8	501	100
pdhC	4	480	100
pgm	6	450	100

Figure 17. Sequence typing allele results window.

3.3.2 Antimicrobial resistance gene detection

The gene detection analysis follows the same workflow as sequence typing. After selecting the AMR scheme, the user can choose one reference database for antimicrobial resistance gene detection. Currently, WebSTing has four databases available: the Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT), the Comprehensive Antimicrobial Database (CARD), MEGARes, and the NCBI's Bacterial Antimicrobial Resistance Reference Gene Database[15] (Gupta, et al., 2014; Jia, et al., 2017; Lakin, et al., 2017). Read samples are provided by specifying the corresponding SRA accession numbers or by direct upload of read files.

[15] https://www.ncbi.nlm.nih.gov/bioproject/313047

3.3.3 Phylogenetic analysis

Phylogenetic analysis through WebSTing shows evolutionary relationships among samples. Samples, previously characterized by MLST or cgMLST analysis, can be selected for comparison through a phylogenetic analysis that relies on locus-based distances between samples. First, the user retrieves the samples previously analyzed by selecting the scheme (MLST or cgMLST) and the organism. Then, the user selects the samples to be compared and starts the analysis. Finally, a dendrogram (phylogenetic tree) visually shows the evolutionary relationships between the samples selected (Figure 18).

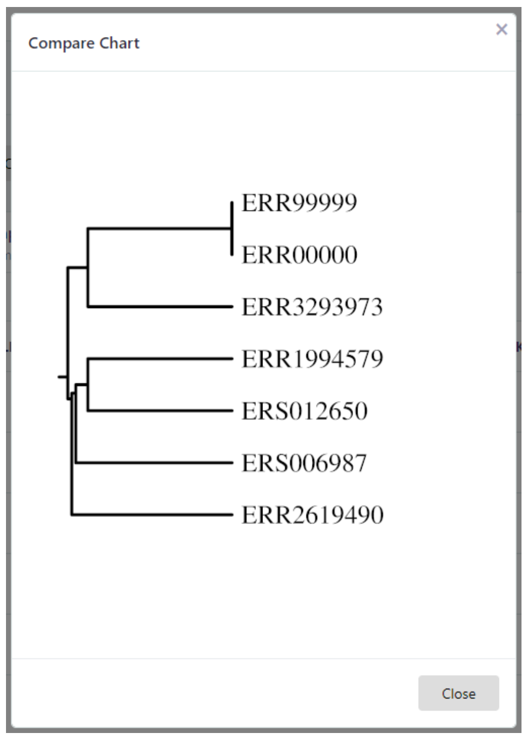


Figure 18. Phylogenetic tree of characterized pathogen samples. WebSTing generates dendrograms from phylogenetic analysis of previously characterized samples.

CHAPTER 4. APPLICATION OF THE STING ALGORITHM TO PUBLIC HEALTH AND ENVIRONMENTAL GENOMICS

4.1 Applying STing to public health: Shiga toxin-producing *Escherichia coli* (*E. coli*) virulence profiling

Shiga toxin-producing *E. coli* (STEC), also called verocytotoxigenic *E. coli* (VTEC), is the denomination for *E. coli* strains that produce toxins similar to the Shiga toxins 1 (Stx1) and 2 (Stx2) produced by *Shigella dysenteriae*. Toxins produced by STEC cause diarrheal illness that include symptoms such as abdominal cramps, fever, vomiting, and bloody diarrhea. Severe cases of STEC infections may derive to the hemolytic uremic syndrome (HUS), a life-threatening blood disorder that causes acute kidney failure. It is estimated that 265,000 STEC infections occur yearly in the US, causing around 3,600 hospitalizations and 30 deaths[16] (Gould, et al., 2009). Although people of all ages may become infected from STEC, young children (5 years old and younger) and elderly (65 and older) have more risk of infection and are more likely to develop HUS[17]. STEC infection can occur through the ingestion of contaminated water or food (undercooked beef, and both raw produce and milk), contact with animals, and contact with infected people. The high number of infections and the risk of developing HUS make STEC a significant concern in public health.

Rapid and accurate STEC infection diagnosis are critical factors for ensuring proper treatment of illnesses and control of outbreaks. Conventional methods for diagnosis of

[16] https://www.cdc.gov/ecoli/pdfs/CDC-E.-coli-Factsheet.pdf
[17] https://www.cdc.gov/ecoli/general/index.html

STEC include the detection of the virulence factor genes Shiga toxin 1 and 2 (*stx1* and *stx2*, respectively), intimin (*eae*), and enterohemolysin A (*ehxA*), which are clinically relevant since they are associated with disease severity. Traditional wet lab methods for detecting and characterizing STEC include culture assays, and non-culture assays such as enzyme immunoassays, cell cytotoxicity, and the Polymerase Chain Reaction (PCR), the gold-standard technique currently used in public health laboratories (Gould, et al., 2009; Parsons, et al., 2016).

Current advances in NGS technologies have led to new genome-enabled techniques for characterizing STEC using WGS data. Raw reads from sequenced STEC isolates by WGS are assembled to produce complete genomes that used for detecting and characterizing genes of interest by using sequence alignment. Although these genome-enabled techniques are less time- and labor-consuming than traditional wet-lab methods, they may not be efficient enough for analyzing the amount of data produced by NGS technologies at the speed required for real-time outbreak control. New rapid and accurate methods are required for gene characterization of STEC isolates using NGS data. This section presents an application of the STing algorithm to the virulence profiling of STEC isolates using WGS.

4.1.1 Materials and methods

4.1.1.1 Dataset

A total of 5,000 whole genome sequencing samples of Enterobactericeae strains isolates from five species (Table 9) were used for this experiment. The dataset corresponds to isolates sequenced between 2014 and 2017 by the Enteric Diseases Laboratory Branch (EDLB) at the CDC, through its network PulseNet and affiliated laboratories. The dataset is publicly available on the NCBI website under the bioproject PRJNA218110.

Table 9. Number of samples per species used in the STEC virulence profiling.

Species	Number of Samples
Escherichia coli	4,989
Escherichia albertii	5
Shigella dysenteriae	2
Shigella sonnei	2
Shigella boydii	1

4.1.1.2 Virulence gene allele database

The database used for this experiment comprises 201 allele sequences of four virulence factor genes (Table 10), obtained from the *E. coli* database (Updated 16 March 2016) (Joensen, et al., 2014) of the VirulenceFinder application from the Center for Genomic Epidemiology[18].

[18] https://cge.cbs.dtu.dk/services/data.php

Table 10. Number of sequences per gene in the database used in the STEC virulence profiling.

Gene Name	Number of Sequences
stx1	23
stx2	121
eae	45
ehxA	12

4.1.1.3 Assembly-based gene detection and novel allele identification

Each isolate sample was assembled independently using two *de novo* assemblers with default parameters: SPAdes (Bankevich, et al., 2012) (v3.10.1), and ABySS (v2.0.2). Then, the program BLASTn (v2.2.28+) (Camacho, et al., 2009) was used for searching the sequences from the virulence database on each assembled genome for detecting candidate alleles to be present in the samples. Alleles were identified as the BLAST hits with query coverage and an identity of 90% from both SPades and ABySS assemblies. Additionally, potential novel alleles were identified as the BLAST hits with query coverage of 100% and one or more base pairs mismatches. The final set of detected alleles and novel alleles was defined as the union of the alleles recovered from both SPades and ABySS assemblies.

4.1.1.4 PCR-based gene detection

PCR-based characterization of the isolates was previously performed by the EDLB at the CDC, PulseNet, and PulseNet affiliated laboratories, following the recommendations and guidelines specified by the CDC (Gould, et al., 2009). PCR available results from

5,000 isolates for genes *stx1* and *stx2*, and from 2,830 isolates for genes *eae* and *ehxA* were used in this experiment.

4.1.1.5 STing-based gene detection

A database index was constructed from the virulence database using the STing indexer program (v0.24.2). Then, the STing typer utility (v0.24.2) with $k = 30$ (*k*-mer size) was used along with the index for searching the *stx1* and *stx2* alleles from the database in all the 5,000 WGS isolate samples, and for searching the *eae* and *ehxA* alleles in the 2,830 WGS samples with available PCR results. From the information reported by STing after the sequence typing analysis, two parameters were defined to predict the presence/absence of the *stx1* and *stx2* genes in the samples: (1) allele coverage (c), defined as the percentage of the allele length that is covered by at least one *k*-mer match, and (2) *k*-mer fraction (f), defined as the ratio between the minimum and maximum average *k*-mer depth of the two Stx alleles:

$$f = \frac{\min\left(\bar{d}_{stx1}, \bar{d}_{stx2}\right)}{\max\left(\bar{d}_{stx1}, \bar{d}_{stx2}\right)}$$

where $\bar{d}_{stx1}$ and $\bar{d}_{stx2}$ are the average *k*-mer depth of the called alleles for the *stx1* and *stx2* genes, respectively.

Performance in terms of computational resources consumed for detecting the virulence genes was measured in terms of runtime in seconds and the maximum peak of RAM in megabytes required for analyzing each sample.

4.1.1.6 Evaluation of the detection performance

Presence/absence results from the assembly-based detection were used as ground truth for comparing the performance of the STing and PCR methods in terms of correctly detected virulence alleles in the dataset. To assess the performance of detection of each method, the Matthews Correlation Coefficient (MCC) metric was used, which is defined as follows:

$$MCC = \frac{(TP \times TN) + (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where, $TP =$ true positives, $TN =$ true negatives, $FP =$ false positives, and $FN =$ false negatives.

4.1.1.7 STing detection parameter optimization

The parameter optimization process was designed to select the minimum values for the allele coverage and *k*-mer fraction parameters for which the MCC is maximized for detecting the virulence genes. A grid of different values for the allele coverage (1 to 100 on increments of 1) and *k*-mer fraction (0.0 to 1.0 on increments of 0.001) parameters were generated. The STing typer results were evaluated in terms of the MCC obtained with each value combination of the two parameters. The algorithm of the implemented grid search approach is shown in Figure 19.

Algorithm 3: Parameter Optimization

Input :

Genes $G = \{g_1, g_2, \ldots, g_n\}$, n is the total number of genes.

Results R, the table with STing detection results for the set G.

Output:

$P = \{t_1, t_2, \ldots, t_m\}$; t_i is the tuple for the gene i with the best parameter values defined as $t_i = (c_i, f_i, A_i, M_i)$, c_i is the allele coverage, f_i is the k-mer fraction, and $M_{i,j,l}$ is the Matthews Correlation Coefficient.

```
function ParameterOptimization(G, R)
    T ← {}
    for each g ∈ G do
        c ← 1
        while c ≤ 100 do
            f ← 0.1
            while f ≤ 1.0 do
                (TP, FP, TN, FN) ← EvaluateResults(R, g, c, f)
                M ← ComputeMCC(TP, FP, TN, FN)
                T ← T ∪ {(g, c, f, M)}
                f ← f + 0.1
            c ← c + 1
    P ← {}
    for each g ∈ G do
        (c′, f′, M) ← SelectBestParameters(T, g)
        P ← P ∪ (c′, f′, M)
    return P
```

Figure 19. Detailed algorithm of the STing parameter optimization for detecting the *stx1* and *stx2* genes. The input of the algorithm includes the names of the genes from which the optimal parameters will be searched and the results from the STing typer utility. The algorithm output is a table with the best values for the allele coverage and *k*-mer fraction parameters for each gene and their corresponding Accuracy and MCC. Best values are selected as the minimum that maximize the MCC.

4.1.2 Results and discussion

Three detection approaches were used in this experiment to assess the ability of STing to detect alleles of the four virulence genes *stx1*, *stx2*, *eae*, and *ehxA* in 5,000 STEC WGS samples: (1) assembly-based, (2) PCR, and (3) STing (Figure 20).

The assembly approach was conceived to establish the presence/absence of the virulence genes in the dataset for being used as ground truth to evaluate the detection performance of the other two approaches. For this purpose, the 5,000 WGS samples from the isolates were assembled into genomes, and alleles of the four virulence genes were searched in the assemblies. One consideration in the assembly-based approach was repetitive regions of the *stx1* and *stx2* genes, which make it difficult to reconstruct the full gene sequence during the de *novo assembly* of the WGS reads. To increase the chances of recovering full sequences for these genes (Figure 20, green), two *de novo* assemblers, SPAdes and ABySS, were used to assembly each WGS read sample. A database (allele DB) with allele sequences of the genes *stx1*, *stx2*, *eae*, and *ehxA*, was used for establishing the presence/absence of the virulence genes and identifying novel alleles on the assembled genomes by using local sequence alignment (BLAST). 14,099 alleles from the four genes were identified in the assembled genomes with the two assemblers (Figure 21), from which 5,375 were novel sequences (Table 11) that were added to the allele DB to create an extended version (extended DB).

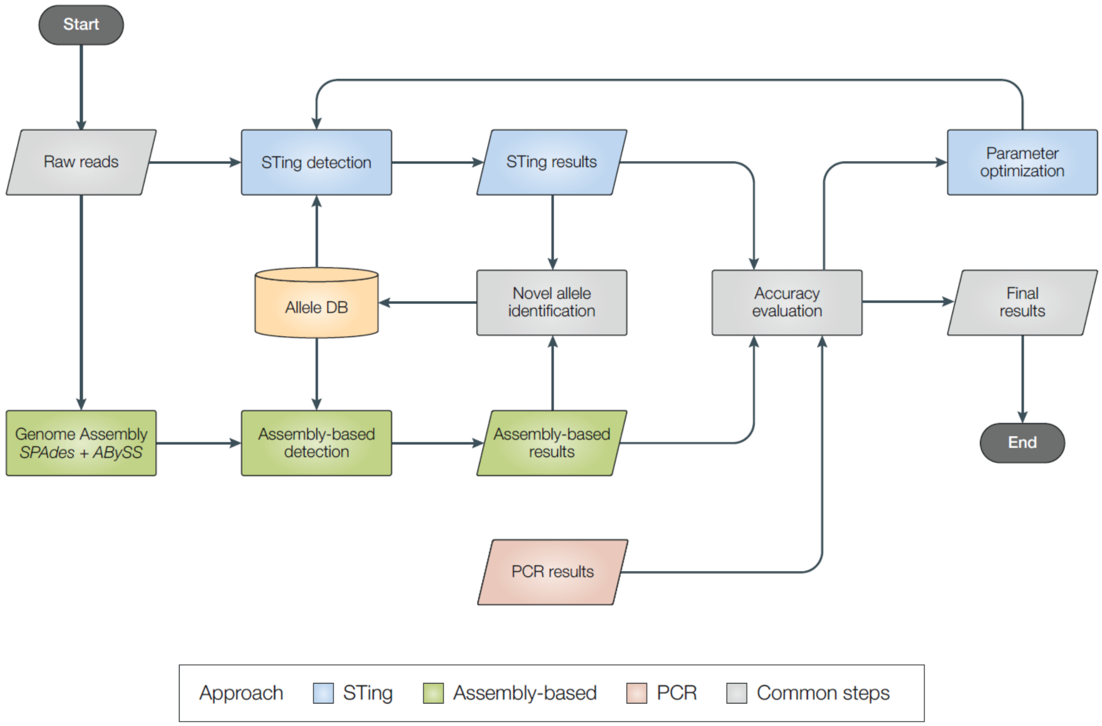


Figure 20. General workflow for the STEC characterization study. Three approaches were used for detecting the virulence factor genes from the Allele DB in the STEC samples. (1) Assembly-based (green), whose results were used as ground truth for evaluating the accuracy of the other two approaches, (2) STing (blue), and (3) PCR (red).

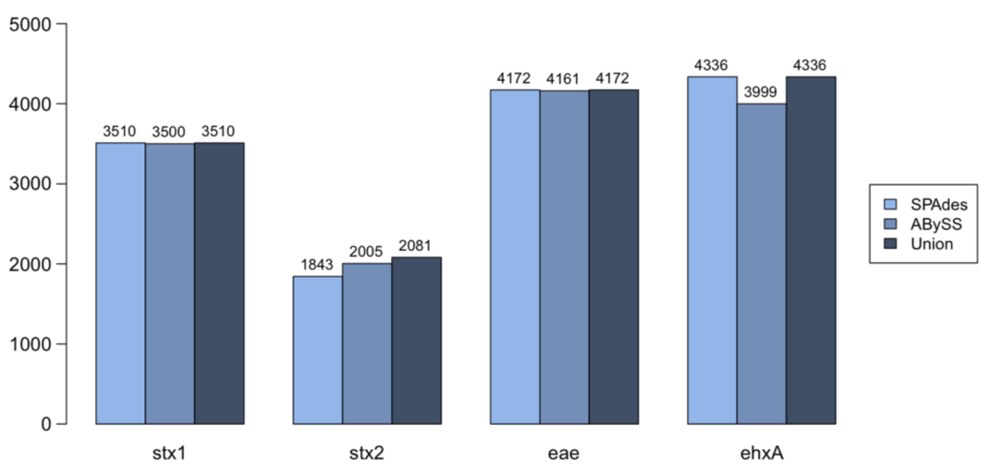


Figure 21. Number of alleles identified in the genomes assembled with SPAdes and ABySS.

Table 11. Number of novel alleles identified from the assembled genomes.

Gene	Number of novel alleles
stx1	266
stx2	760
eae	2458
ehxA	1891

Results from previous PCR characterization of the four virulence genes in the 5,000 STEC isolates were used in the PCR approach (Figure 20, red). These results were obtained from public health laboratories and evaluated using the presence/absence results derived from the assembly approach to measuring the performance of the PCR method for detecting the four virulence genes.

In the STing approach (Figure 20, blue), the typer utility was used for detecting the virulence genes from the allele DB by processing the raw read samples. A first evaluation of the results showed a performance in terms of the MCC between 0.78 and 0.90 for detecting the four genes (Figure 22). This relatively poor performance is due to false positives and false negatives predictions. False positives are likely due to *k*-mers that match small portions of the target genes, but that are derived originally from reads that correspond to regions of the genome different from the virulence genes. Although these matches do not cover the target gene entirely, they are considered by STing when the alleles are called following the *k*-mer-based strategy of choosing the sequence with the highest count of *k*-mer matches. In these cases, STing reports the call with a "*" character to indicate that the allele has a coverage below 100%. As an example, Figure 23 shows the *k*-mer depth distribution of a false positive result for the *stx2* gene (stx2_27), and a true positive for the *stx1* gene (stx1_12) in one of the STEC samples analyzed. The allele stx1_12 is fully covered by *k*-mer matches and it has a high average *k*-mer depth in comparison with the allele stx2_27, which is partially covered and has a low *k*-mer depth. False negative results are likely due to alleles present in the samples that are not present in the Allele DB.

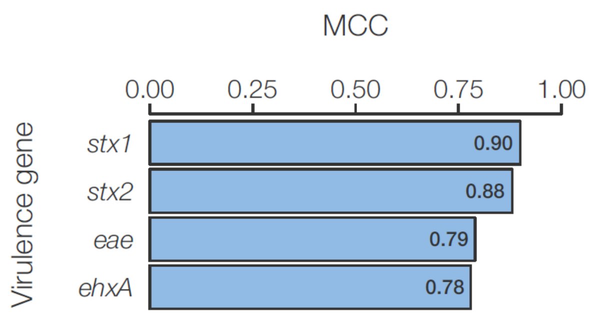


Figure 22. Virulence gene detection performance of STing. The plot shows the performance of detection of STing in terms of the Matthews correlation coefficient (MCC) of STing using default parameters.

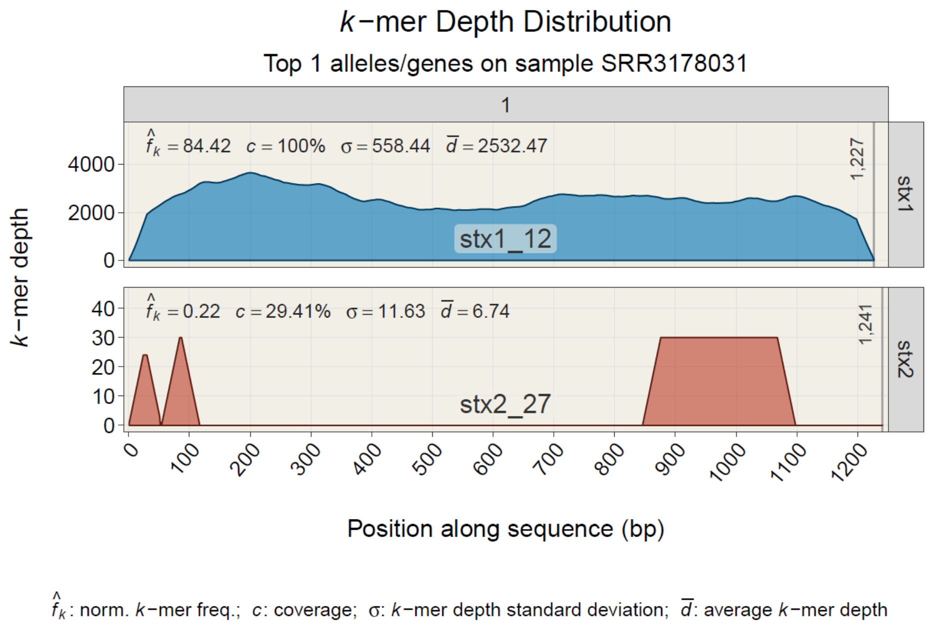


Figure 23. *k*-mer depth distribution of a true positive and a false positive predicted allele with STing. The plot shows the *k*-mer depth distribution along the sequence of the predicted allele 12 for the gene *stx1* (stx1_12, blue), and the predicted allele 27 for the gene *stx2* (stx2_27, red). The distribution of the allele stx1_12 is an example of a true positive prediction in which complete sequence is covered ($c = 100\%$) with a high average *k*-mer depth ($\bar{d} = 2532.47$). However, the allele stx2_27 is covered partially ($c = 29.41\%$) with a low average *k*-mer depth ($\bar{d} = 6.74$), which is indicative of a false positive prediction.

To improve the detection performance of STing, it was necessary to reduce the number of false positive and false negative predictions. To reduce the false negatives, the presence/absence of a gene was redefined according to two parameters derived from the results of the typer utility: (1) the allele coverage (*c*), and (2) the *k*-mer fraction (*f*). The allele coverage is the percentage of the predicted allele that is covered by *k*-mer matches, and it is applied for the detection of the four virulence genes. The *k*-mer fraction is defined as the ratio between the minimum and maximum average *k*-mer depth of the *stx1* and *stx2* genes, and it is applied only for the detection of the Stx genes. The optimal values for these two parameters were defined by using a grid search-based optimization process to find the minimum values for *c* and *f* that maximize the number of genes correctly predicted in the samples (Figure 24). Figure 25 shows a heatmap as a result of the parameter optimization process for the gene *stx1*, and Table 12 shows the optimal parameter values for each virulence gene. To reduce the false negatives, the extended DB containing the novel alleles identified through the assembly approach was used to detect the virulence genes in the STEC samples. STing results after implementing the *c* and *f* parameters and using the extended DB showed an improvement between 10% and 20% in performance measured as the MCC when detecting the four virulence genes (Figure 26).

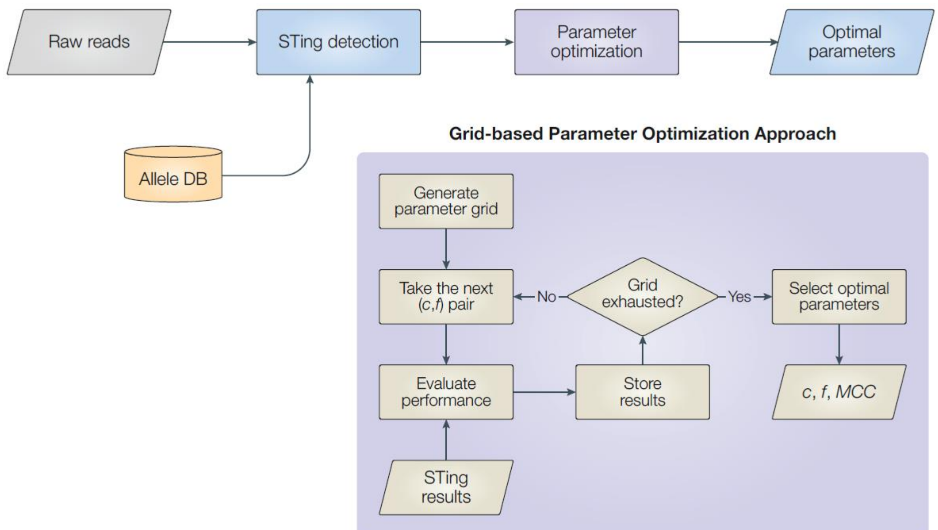


Figure 24. Schematic representation of the grid-based parameter optimization process. A grid of values for the allele coverage (*c*) and *k*-mer fraction (*f*) parameters was explored to evaluate the STing detection results. For each combination of *c* and *f*, the performance of STing was evaluated using the Mathews Correlation Coefficient (*MCC*) metric. In the end, the minimum values of *c* and *f* that result in the best detection performance, *i.e.*, maximum *MCC*, are selected as optimal parameters.

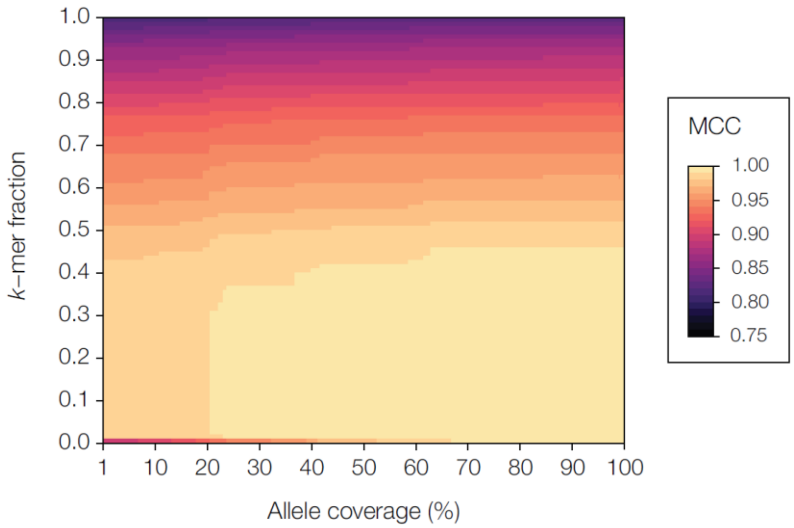


Figure 25. Heatmap of the grid-based parameter optimization process for the gene *stx1*. The plot shows the MCC achieved with STing when detecting the *stx1* gene using different values for the allele coverage (*c*) and the *k*-mer fraction (*f*) parameters. Lighter colors indicate better gene detection performance.

Table 12. Optimal values for *c* and *f* for detecting each virulence gene.

Gene	*c*	*f*
stx1	84.48	0.02
stx2	97.60	0.02
eae	96.10	-
ehxA	97.50	-

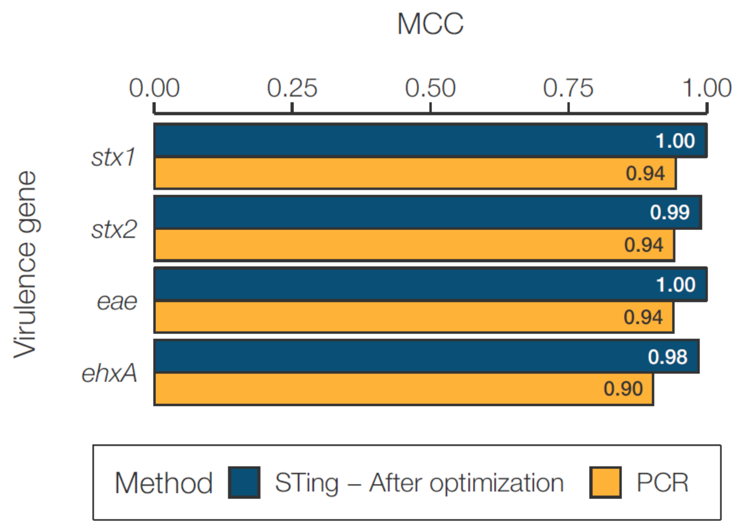


Figure 26. Virulence gene detection performance of STing and the PCR method. Comparison of the performance measured as the MCC of STing and the PCR method on detecting each virulence gene. The STing's performance showed correspond to the results after optimizing the detection process by using the *c* and *f* parameters, and the extended DB.

The comparison between STing and PCR results (Figure 26) shows that STing has better performance in detecting all the four virulence genes in the STEC samples analyzed. STing was 100% accurate (MCC = 1.0) on detecting the *stx1* and *eae* genes and showed marginal drop-off for the *stx2* and *ehxA* genes.

Performance in terms of computational resources showed that STing is fast and has a light memory footprint, requiring an average of 31.33 seconds and 44.80 MB of RAM for analyzing each sample (Table 13). The low consumption of resources enables the use of STing for virulence profiling in most common personal computers including laptops.

Table 13. Performance of STing in terms of computational resources for detecting the virulence genes.

Statistic	Runtime (s)	RAM peak (MB)
Min	1.93	43.10
Max	191.66	106.80
Average	31.33	44.80

Results showed that STing is a better technique for characterizing virulence genes in STEC samples compared to the PCR method. Although PCR is the gold-standard method used in public laboratories for isolate characterization, it fails to detect new alleles of the virulence genes in STEC samples. The PCR method relies on specific primers designed to capture regions from the DNA of an isolate that corresponds to the genes of interest. Design of such primers is based on a set of alleles previously characterized. However, as the target genes change due to horizontal gene transfer events in bacterial organisms, the ability of the PCR primers to capture new alleles is compromised. Moreover, PCR is a time-consuming and labor-demanding technique, features that are not desirable in real-time molecular epidemiology. On the contrary, STing is a lightweight and fast software solution that provides more accurate results and has the ability to detect novel alleles in STEC samples. Potential novel sequences are reported by STing as alleles with coverage below 100% and marked with an "*" character. In such cases, the reported sequence corresponds to the most similar allele in the index database. This feature enables the recovering of the novel allele by using *de novo* assembly of the corresponding WGS sample.

4.1 Applying STing to environmental genomics: *nifH* gene-based taxonomic assignment of amplicon sequencing samples

The *nifH* gene encodes for the dinitrogenase reductase subunit, an enzyme of the nitrogenase complex involved in the fixation of atmospheric nitrogen into other forms (*e.g.*, ammonia) that can be metabolized by living organisms. This process is known as biological nitrogen fixation. Bacteria or archaea capable of fixing atmospheric nitrogen, referred to as diazotrophs, are crucial for life since they are the only natural source for bioavailable nitrogen on the earth.

The *nifH* gene is the biomarker most widely used by the microbial ecology research community to characterize nitrogen-fixing organisms, including bacterial and archaeal species. In addition to metagenomics studies, amplicon sequencing studies are a standard method for characterizing taxonomic markers (16S RNA), or biochemical markers (*nifH* genes). *nifH* amplicon sequencing studies, for example, are intended to characterize the capacity for biological nitrogen fixation in a specific environment. To perform such characterization, *nifH* amplicon sequences are compared to a reference database of *nifH* sequences from a diverse range of known bacterial and archaeal species.

Traditional methods for amplicon sequencing-based biomarker characterization usually depend exclusively on sequence alignment software like BLAST for sequence comparison. For example, TaxADivA[19] is a pipeline that uses BLAST for taxonomy assignment and diversity assessment of *nifH* amplicon sequences (Gaby, et al., 2018). Recently, the environmental microbial research community is increasingly relying on NGS

[19] https://github.com/lavanyarishishwar/taxadiva

technologies to generate amplicon sequences for characterization studies. The amount of data produced by these innovative technologies poses a computational challenge for data analysis based on sequence alignment. This section explores the implementation of the alignment-free STing algorithm to classify *nifH* amplicon sequencing reads, as an alternative method for the efficient characterization of nitrogen-fixing organisms.

4.1.1 Materials and methods

4.1.1.1 Algorithm description

Given an amplicon sequencing read sample, STing can classify each read according to a reference gene database with known taxonomic information. STing classifies the reads by using exact *k*-mer matching and a frequency counting strategy, as the algorithms for sequence typing and gene detection previously described in this work. As in the sequence typing and gene detection algorithms, STing *k*-merizes each read and count the matches of each *k*-mer in the database. However, unlike sequence typing and gene detection, the selection of the best sequence occurs at the read level, *i.e.*, after searching the *k*-mers extracted from each sequence read and not at the sample level. This selection strategy allows for classifying each read by assigning the taxonomy of the gene with the highest count of *k*-mer matches.

The STing algorithm for read classification comprises two main phases: (1) database indexing, and (2) read classification (Figure 27). The following sections describe each phase:

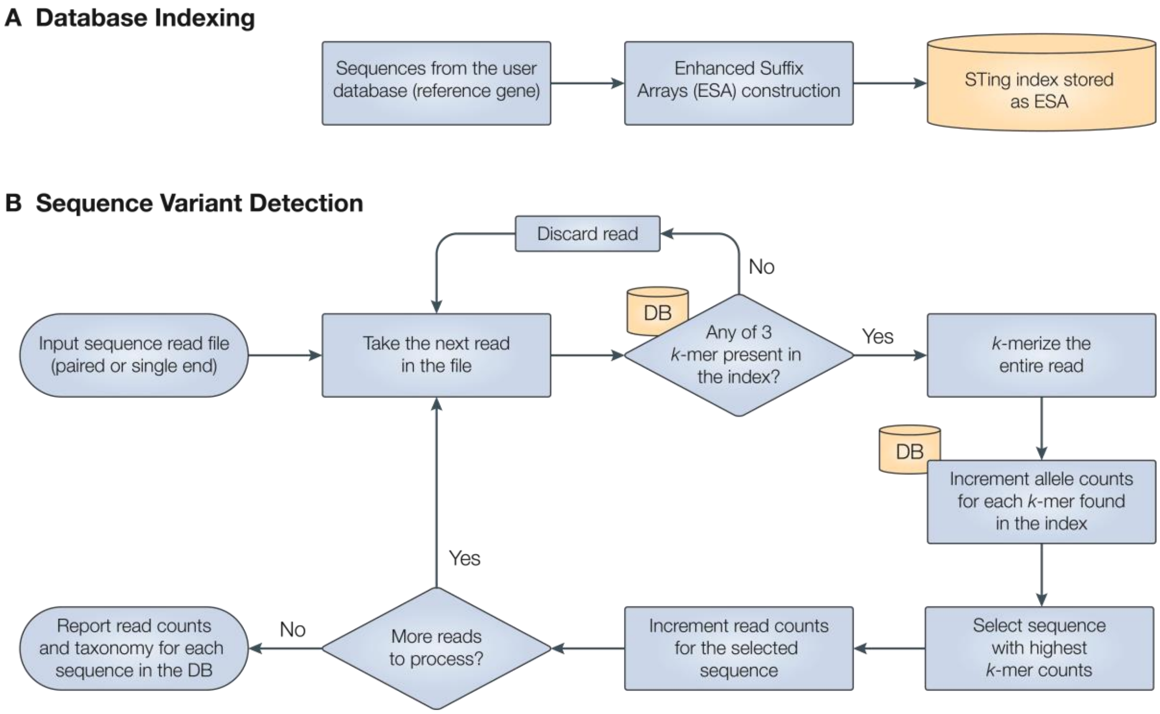


Figure 27. A detailed flowchart of the STing algorithm for reads classification. The read classification algorithm comprises two main phases: (**A**) Database indexing–user-supplied reference gene sequences are transformed into an Enhanced Suffix Array (ESA) index database for a rapid *k*-mer search for read classification, and (**B**) Sequence variant detection. The *k*-mers located at the beginning, middle, and end of each read are searched in the ESA index. If there are no matches, the read is discarded; otherwise, the read pass to the next step. Reads that passed the filter are fully *k*-merized, and each *k*-mer is searched in the ESA index. For each match in the index, a table of frequencies (*k*-mer frequency table) is updated for the matched genes. Then, each *k*-mer frequency is normalized by dividing the count by the length of the corresponding gene. The gene with the highest normalized *k*-mer frequency is selected as the best sequence. The read is classified by assigning the taxonomy associated with the best gene, and a table of read counts is updated for the best gene. This process repeats until all the reads of the input dataset have been processed. Finally, the genes from the database with read counts, *i.e.*, with assigned reads, are reported with their corresponding read count and associated taxonomy.

4.1.1.2 <u>Database indexing</u>

In this phase, executed though the STing indexer utility, the algorithm constructs an ESA index used during the read classification (Figure 27A). The indexer requires a multi-fasta file with all the sequences of the reference genes and an additional profile file that contains the associated taxonomy of each sequence. Although the profile file can have any level of taxonomic annotation, the more extensive the taxonomic information for each sequence, the better the final classification of the reads. A complete lineage for each sequence should have taxa for seven taxonomic ranks: kingdom, phylum, order, class, family, genus, and species. The indexer constructs a single ESA index of all the gene sequences provided (gene index) and stores the profile table.

4.1.1.3 Read classification

The STing classifier utility is used for the read classification phase. In this phase, the algorithm classifies or assigns a read from the input dataset to a gene in the database index. The read classification comprises four algorithmic steps: (1) read filtering, (2) *k*-mer counting, (3) read assignment, and (4) reporting (Figure 28).

In the read filtering step, the algorithm searches the *k*-mers located at the beginning, middle, and end of each sequence within the gene index database. Let l be the length of a read. The start position of each of the three *k*-mers is defined as follows: $s_b = 1$, for the initial *k*-mer; $s_m = \left\lfloor \frac{l-k}{2} \right\rfloor$, for the middle *k*-mer; and $s_m = l - k + 1$, for the *k*-mer at the end. If none of the three *k*-mer is found in the allele index, the algorithm discards the read; otherwise STing passes the read to the next step. The default size is k=30, and users can change this value.

Algorithm 3: STing Read Clasification

Input :

Genes of reference $G = \{g_1, g_2, \ldots, g_m\}$, where m is the total number of reference genes.

Reads $R = \{r_1, r_2, \ldots, r_n\}$, n is the total number of reads.

k-mer size $k \leq \min(\text{length}(R))$.

Gene index $\mathcal{G}$, the generalized ESA index of G.

Taxonomy table $T = \{l_i \mid 1 \leq i \leq m\}$, where l_i is the lineage of the gene g_i defined as $l_i = (t_1, t_2, \ldots, t_o)$, a o-tuple of taxa t_j.

Output:

Read counts table $C = \{c_i \mid 1 \leq i \leq m\}$, where $c_i = (g_i, f_i, l_i)$, a 3-tuple that contains the count of assigned reads f_i for the gene g_i and its corresponding lineage l_i.

procedure READCLASSIFICATION($G, R, k, \mathcal{G}, T$)
- $F \leftarrow []$ ▷ Read counts
- **for each** $r \in R$ **do** ▷ Read processing
  - $(first\_kmer, mid\_kmer, last\_kmer) \leftarrow$ GETFIRSTMIDANDLASTKMER(k, r) ▷ **(i)** Filtering
  - **if** $first\_kmer \notin \mathcal{A} \wedge mid\_kmer \notin \mathcal{A} \wedge last\_kmer \notin \mathcal{A}$ **then**
    - **continue**
  - $freqs \leftarrow []$ ▷ **(ii)** k-mer counting
  - $K \leftarrow$ GETALLKMERS(k, r)
  - **for each** $kmer \in K$ **do**
    - $matched\_gene \leftarrow$ FIND($kmer, \mathcal{G}$)
    - **if** $matched\_gene \neq \emptyset$ **then**
      - $freqs[matched\_gene] \leftarrow freqs[matched\_gene] + 1$
  - $norm\_freqs \leftarrow$ NORMALIZEFREQS($freqs, \mathcal{G}$) ▷ Normalize by gene length
  - $best\_gene \leftarrow \arg\max(norm\_freqs)$ ▷ **(iii)** Classify read
  - $F[best\_gene] = F[best\_gene] + 1$
- **for each** $i \in read\_counts$ **do** ▷ **(iv)** Reporting
  - **print** $G[i] + F[i] + T[i]$

Figure 28. Detailed STing read classification algorithm. Input for the read classification algorithm includes the sequencing reads to be processed, the k-mer size, the list of reference genes, and the database that comprises the gene ESA index, and the profile table with the taxonomic information of each sequence. The output is a read counts table that includes the name of each gene that had assigned reads, the count of read assigned per each gene, and the associated taxonomic information.

In the k-mer counting step, STing k-merizes each read that passed the filter matching stage and then searches each k-mer from the read against the gene sequence index. For each k-mer match in the gene index, the classifier increments a k-mer counter for the matched genes. Then, the algorithm normalizes the *k*-mer frequencies by the length of the corresponding genes.

In the read assignment step, the algorithm selects the gene that has the maximum normalized *k*-mer frequency and assigns the read to this gene. The algorithm increments a counter of reads for the selected gene.

Finally, in the reporting step, the STing reports the name, the number of reads, and the associated taxonomy (from the profile table) for each gene with assigned reads using the Biological Observation Matrix (BIOM) format[20] (McDonald, et al., 2012).

4.1.1.4 *nifH* gene databases

The STing databases for this experiment arose from an initial reference database of 3,693 *nifH* sequences obtained through a fully automated database pipeline for retrieval and validation of *nifH* sequences from the GenBank NRDB/NT database[21]. The pipeline, as well as the database, were developed by Luz Medina-Cordoba, a graduate student from the Kostka/Jordan labs research collaboration, as part of her Ph.D. thesis research for characterization of diazotroph organisms.

[20] http://biom-format.org/
[21] https://www.ncbi.nlm.nih.gov/genbank/

The taxonomic annotation for the initial reference database was retrieved from the NCBI Taxonomy database through a custom script developed for this purpose[22]. The information obtained included seven taxonomic ranks: (super) kingdom, phylum, order, class, family, genus, and species. Sequences (3,412) with taxonomic information for each of the ranks in the taxonomic lineage were retained. Sequences with missing information for one or more ranks were discarded. Then, duplicates were removed from the dataset, which led to a set of 2,886 distinct *nifH* sequences. Two subsets arose from this dataset: (1) 732 full-length sequences ($l \geq 750\ bp$), and (2) 2,154 partial length sequences ($l < 750\ bp$). Each subset and its corresponding taxonomic profile files were processed with the STing indexer utility (v0.24.2) to produce two reference indexed databases for this experiment.

4.1.1.5 Dataset

Four samples were generated from the full and partial length *nifH* sequence subsets (two from each subset). Each sample was generated to represent a different community of nitrogen-fixing organisms. To generate a sample, all genera present in the corresponding sequence subset (full or partial) were randomly sorted and assigned with numbers from 0 to 25 to stand for the relative abundance in a community. Members of a sample were selected by taking the first *n* genera for which their cumulative relative abundance was equal to 100. *nifH* sequences corresponding to the genera selected were chosen randomly from the corresponding subset to build a source sequence file. Finally, seven amplicon

[22] https://github.com/hspitia/binf_scripts/blob/master/get_taxonomic_data.py

Illumina paired-end (MiSeq 2x250 bp) read sets were simulated from each source sequence file with the software ART (v2.5.8) (Huang, et al., 2012). Each read set was simulated at a different sequencing depth: 1, 5, 10, 20, 40, and 100x. The four samples were designated FA, FB (from the full-length subset), PA, and PB (from the partial length subset).

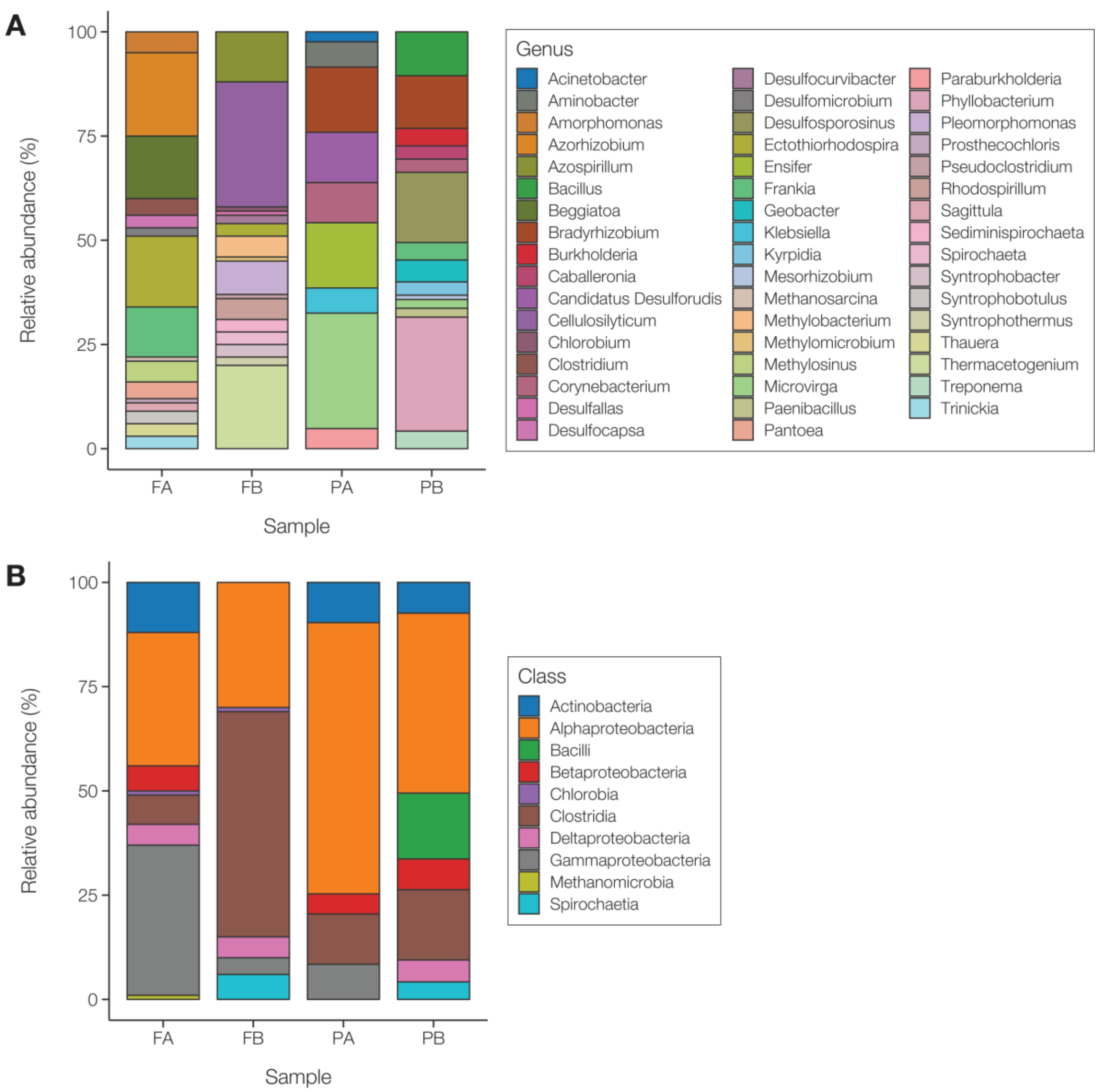


Figure 29. The relative abundance of the samples simulated. The figure shows the relative abundance distribution at the level of (**A**) genus and (**B**) class, for each of the four simulated community samples.

4.1.1.6 Read classification test design

The classifier utility (v0.1) was run on each of the read sets of the four simulated community samples using a $k = 30$. The accuracy of STing for classifying reads at three taxonomic levels (species, genus, and family) from both the classifier results (predicted classification) and the simulated samples (observed classification) involved the calculation of two metrics. The first metric was the relative abundance, defined as the proportion of reads of a given member of the community from the total of reads of the read set analyzed. The second metric was the Shannon diversity index (Spellerberg and Fedor, 2003) which is defined as $H' = -\sum_{i=1}^{n} p_i \ln p_i$, where n is the total number of species, genera, or families in a sample, and p_i is the relative abundance of the i-th species, genus, or family in a sample.

4.1.2 Results and discussion

The *k*-mer frequencies-based algorithm for read classification was implemented in a software utility called classifier, which completes the STing suite of alignment-free programs for NGS data analysis. The STing classifier program was used to classify the sequencing reads from each nitrogen-fixing community simulated sample. The relative abundance and Shannon index were used to evaluate the performance of STing for classifying *nifH* amplicon sequencing reads. The metric values were calculated from the STing classification results (predicted) and then compared to the known metric values of the simulated samples (observed).

In terms of the first metric, the relative abundance, STing achieved better classification ($R^2 = 1$) at the species level (Figure 30A) for the full length-derived samples (FA and FB) than for the partial length-derived samples ($R^2 = 0.71$ for PA, and $R^2 = 0.7$ for PB). At the genus level (Figure 30B), the concordance between the predicted and observed relative abundance was perfect ($R^2 = 1$) for the samples FA, FB, and PA. However, for the sample PB the concordance only reached 50% ($R^2 = 0.51$) at the genus level.

Although it is expected to have an improvement in classification performance for higher taxonomy ranks, *e.g.*, the genus level compared to the species level, the low R^2 value in this case at the higher level is due to the total number of reads that STing reported as unclassified. The differences between predicted and observed counts due to unclassified reads are distributed across all the different members of the simulated community PB at the species level. However, slight differences in read counts of each species are summarized at the genus level, increasing the total difference of counts that affects the classification performance negatively.

As expected, the sequencing depth of the simulated samples did not impact the classification performance significantly (Figure 34 to Figure 37). This result is explained by the strategy of the algorithm for classifying reads. For read classification, the selection of the best sequence from the database to which the read will be assigned is executed at the read level. In other words, the algorithm classifies each read independently based only on its *k*-mer matches.

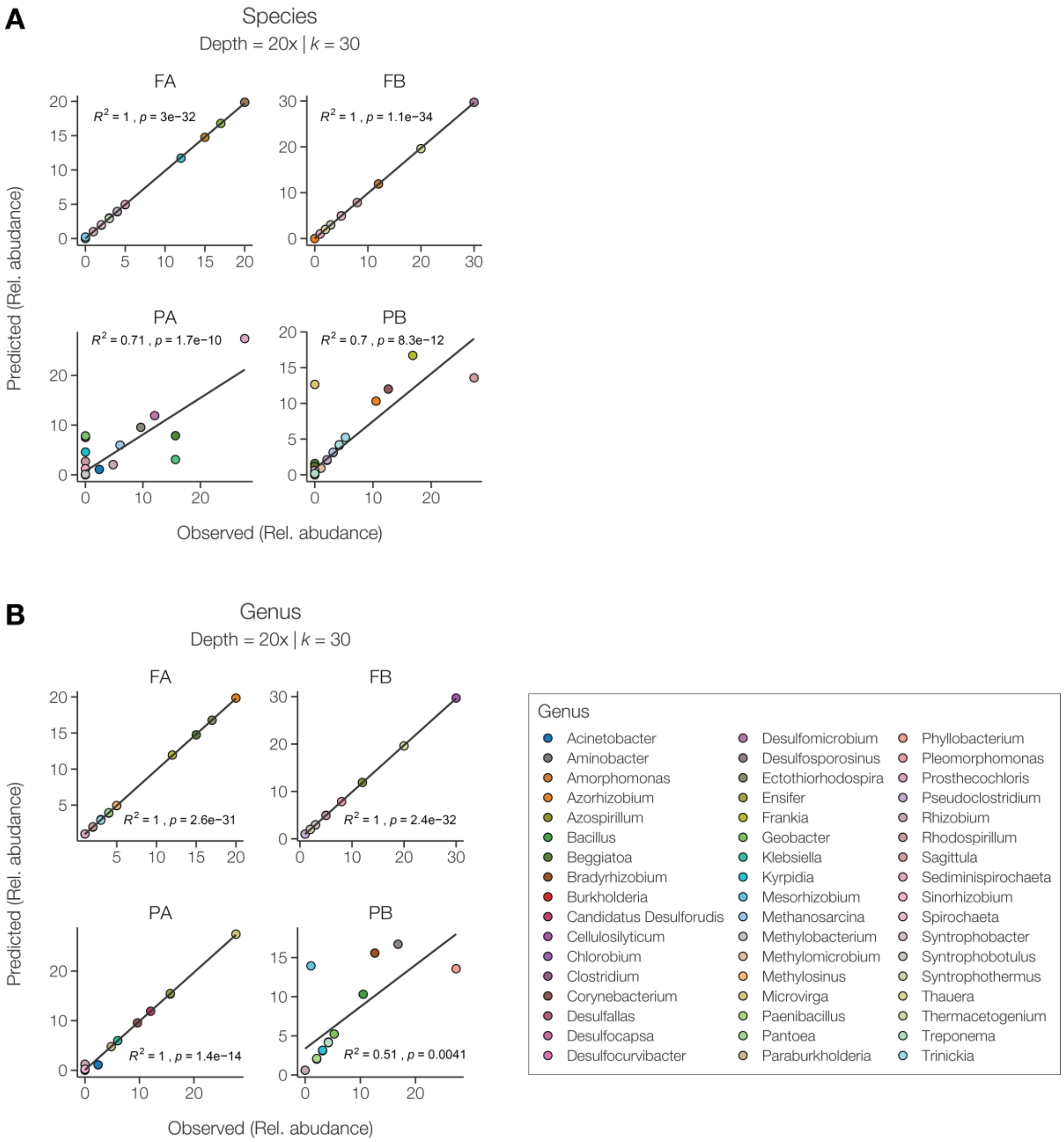


Figure 30. Comparison of the predicted and observed relative abundance. The panels show the relative abundance calculated after classifying the simulated read sets with STing (predicted) as a function of the actual relative abundance of the corresponding simulated dataset (observed), at the level of (**A**) species, and (**B**) genus. Each plot shows the relative abundance calculated from the STing classification using a *k*-mer size of 30, over the read sets simulated at 20x sequencing depth.

The Shannon index, widely used in ecology for comparing diversity between communities, was selected as the second metric to evaluate the STing classification results. This index is based on communication theory and assumes that the members of a community are randomly selected from a large population. The Shannon index considers the richness and evenness of a community and supplies an estimate of biological variability. The overall classification performance of STing across the four simulated samples was evaluated by comparing the predicted and observed diversity measured as the Shannon index. At the level of species (Figure 31), STing had a low performance in classification, reaching only a $R^2 = 0.17$ between the predicted and the observed index values (Figure 31, 3x panel). Here, the effect of the read misclassification is maximized with the Shannon index. However, at higher taxonomic levels, STing had a better classification performance. STing had an R^2 of 0.93 and 1.0 between the predicted and observed diversity at the genus and family levels, respectively.

Classifying *nifH* amplicon sequencing reads is challenging. Sequences from the reference database correspond to the homolog gene *nifH*, thereby sharing a high similarity. To accurately classify reads that came from a genomic region of interest in different organisms, it is necessary to have high-quality reference sequences from distinct species. Chances of having enough variation to distinguish between highly similar gene sequences from distinct species, increase as the reference sequences are close to standing for the complete gene of interest. The results showed in this section confirmed this expectation. STing had better performance on samples that were generated from full-length *nifH* sequences (FA and FB) and classified using the full-length reference database index. Partial length reference sequences may not have enough information to disambiguate the classification between reads that came from highly similar regions in varied species. Suppose two identical partial length reference sequences that correspond to two distinct species and represent a fragment of a very conserved region of the *nifH* gene. During the classification process, when *k*-mers of an input read that comes from one of the two species are searched in the database, it is very probable that both reference sequences have the same *k*-mer frequency. After sorting the normalized *k*-mer frequencies, the sequence that represents the wrong species is in the first place, and the sequence from the correct species is in second place. On even *k*-mer counts, the algorithm will select the sequence that is in the first place, and then the read will be misclassified. One solution to this problem is to assign the read to the two species, *i.e.*, counting 0.5 reads for each species.

Despite the challenges of the read classification problem, the alignment-free approach proposed in this work demonstrates its potential in the environmental genomics field. The STing algorithm avoids all the computationally intensive steps of read quality control, assembly, and sequence alignment, and provides a useful classification performance at the genus or the species level when full-length sequences are used as a reference.

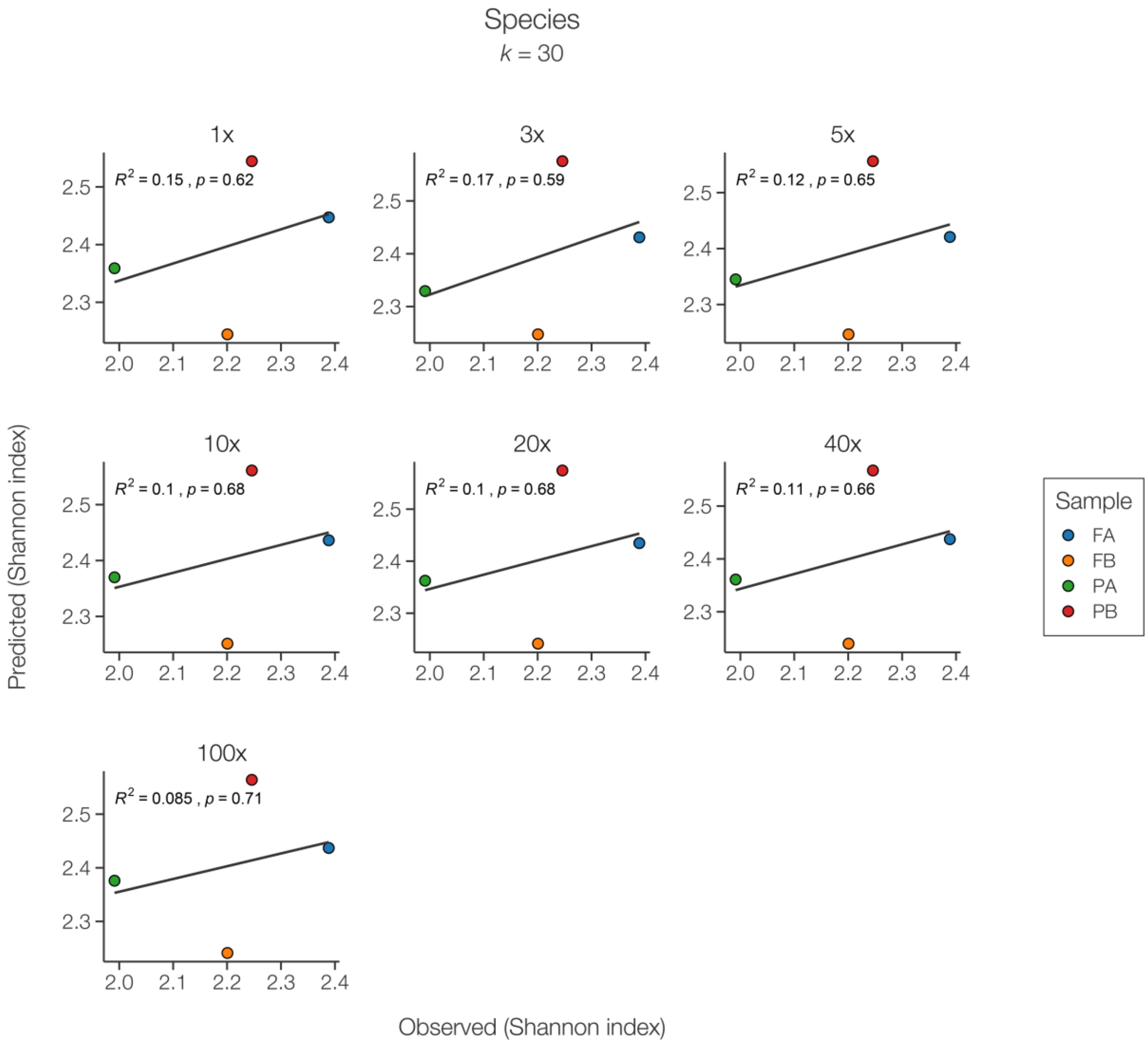


Figure 31. Comparison of the predicted and observed Shannon index at the level of species. The Shannon index calculated after classifying the simulated read sets with STing (predicted) is shown as a function of the Shannon index of the simulated datasets (observed), at the level of species. Each plot shows the Shannon index calculated from the analysis performed with STing using a *k*-mer size of 30, over the simulated samples at a different sequencing depth.

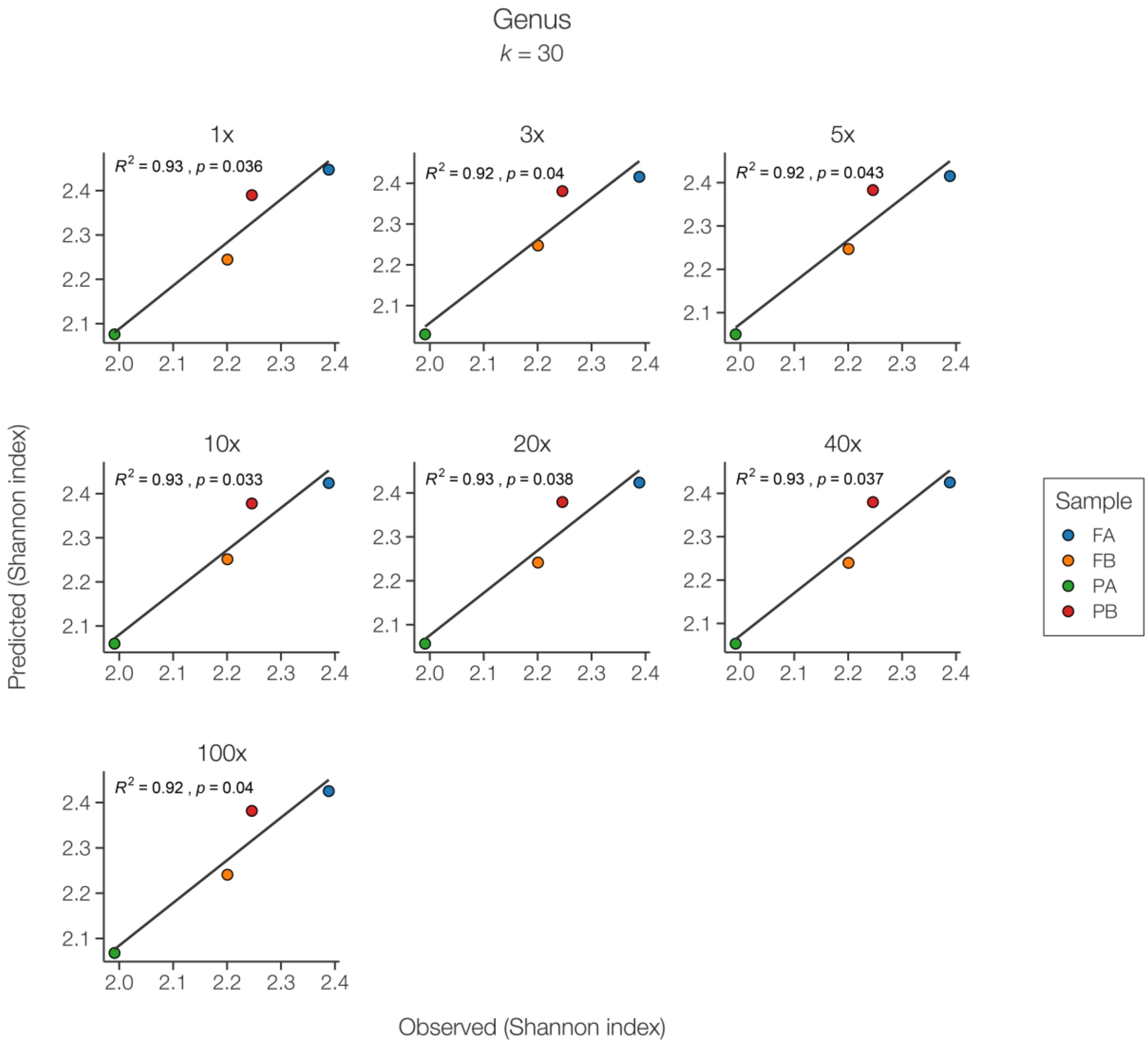


Figure 32. Comparison of the predicted and observed Shannon index at the level of genus. The Shannon index calculated after classifying the simulated read sets with STing (predicted) is shown as a function of the Shannon index of the simulated datasets (observed), at the level of genus. Each plot shows the Shannon index calculated from the analysis performed with STing using a *k*-mer size of 30, over the simulated samples at a different sequencing depth.

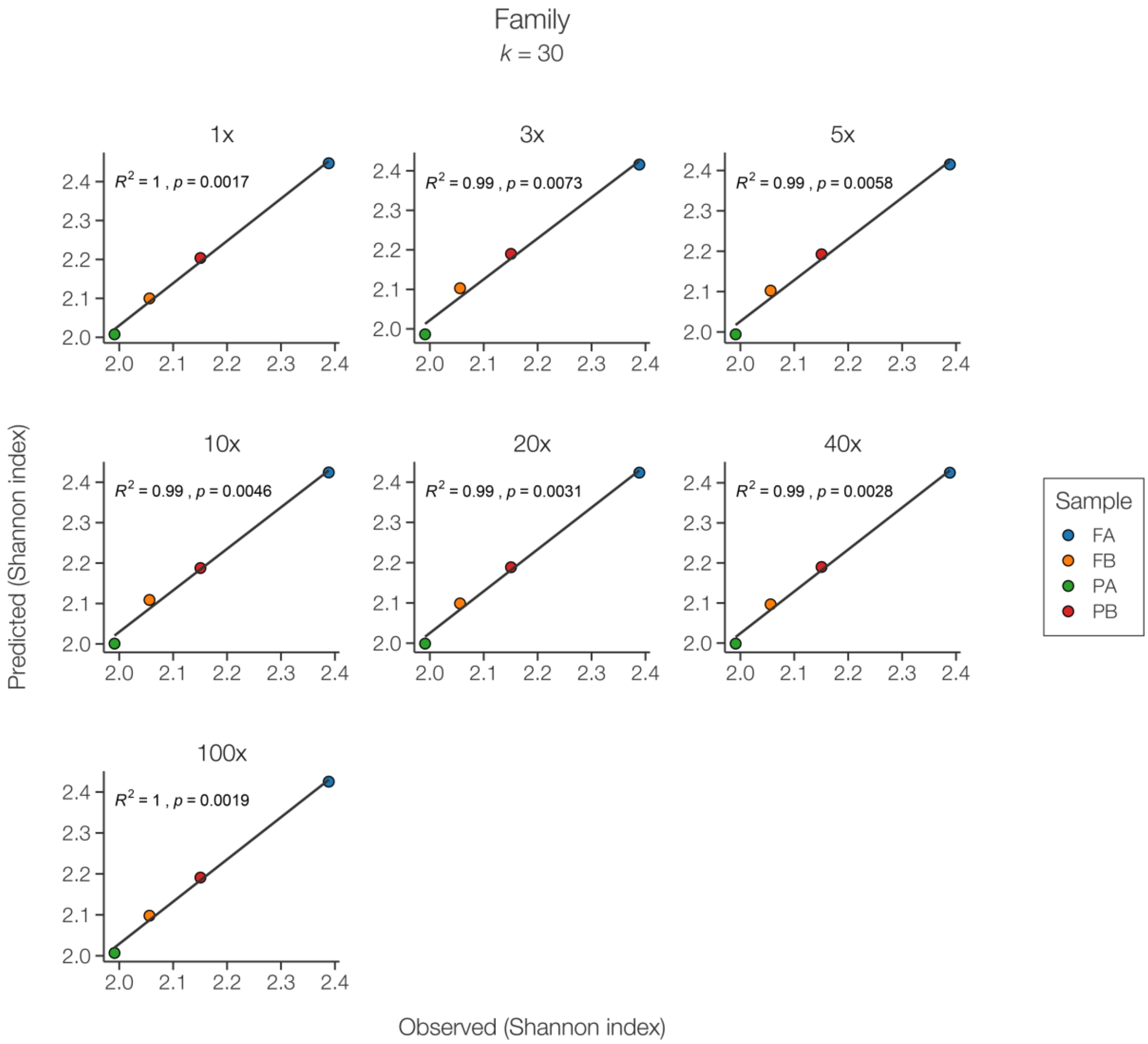


Figure 33. Comparison of the predicted and observed Shannon index at the level of genus. The Shannon index calculated after classifying the simulated read sets with STing (predicted) is shown as a function of the Shannon index of the simulated datasets (observed), at the level of genus. Each plot shows the Shannon index calculated from the analysis performed with STing using a *k*-mer size of 30, over the simulated samples at a different sequencing depth.

CHAPTER 5. CONCLUSIONS AND FUTURE PROSPECTS

Next generation sequencing (NGS) technologies have propelled an unprecedented adoption of genome-enabled approaches to molecular epidemiology. Public health agencies around the world are increasingly relying on whole genome sequencing (WGS) data and genome-enabled bioinformatic methods for surveillance and control of infectious diseases. Despite the advantages that complete genomes provide for molecular epidemiology, there are substantial challenges related to the analysis of the massive amount of data produced by NGS technologies. Traditional genome-enabled bioinformatics methods for molecular epidemiology rely on the computationally expensive and time-demanding tasks of genome assembly and sequence alignment. These tasks currently represent a critical bottleneck for the routine use of genome-enabled approaches in molecular epidemiology. In this thesis research, I developed an assembly- and alignment-free algorithm for efficient analysis of NGS data, which I implemented into a suite of turn-key software applications for NGS-based molecular epidemiology and environmental microbial genomics.

Chapter 2 presented the development and implementation of an alignment-free algorithm – STing – based on a novel exact *k*-mer matching approach for analyzing unprocessed sequencing reads, without executing sequence quality control, genome assembly, and sequence alignment. The STing algorithm was implemented into two turn-key software applications – the typer and detector methods for rapid and accurate genome-enabled characterization of bacterial pathogens. The STing typer utility was compared to six of the most widely used programs for genome-enabled sequence typing, using the traditional MLST scheme, and two larger typing schemes, rMLST, and cgMLST. Results showed that STing outperformed the other sequence typing applications in terms of accuracy and efficiency using the MLST and rMLST schemes and was second with the cgMLST scheme. STing was the only application able to perform the analysis using all three of the typing schemes tested, showing the ability to scale to genome-enabled schemes successfully. Expanding the applicability of the *k*-mer frequencies approach, the STing algorithm was implemented into an NGS-based gene detection utility – STing detector. STing's detector was used to detect two epidemiologically important types of markers: antimicrobial resistance (AMR) genes and virulence factor (VF) genes. Results showed 100% accuracy of STing in detecting AMR and VF genes on NGS samples from species of high priority in clinical microbiology research. STing is currently the only application for genome-enabled bacterial pathogen characterization that supplies assembly- and alignment-free sequence typing and gene detection.

Responding to the need for easy-access tools that provide genome-enabled approaches to molecular epidemiology to public health laboratories that lack bioinformatics infrastructure and human expertise, I developed WebSTing. As presented in Chapter 3, WebSTing is an easy-to-use Web platform for the genome-enabled automated characterization of bacterial pathogens using the STing algorithm. WebSTing provides assembly- and alignment-free sequence typing, gene detection, and phylogenetic analysis of WGS samples of bacterial isolates. WebSTing was developed with state-of-the-art software development and security standards to provide full scalability on cloud environments.

In addition to the STing algorithm development and implementation, I explored the applicability of the STing algorithm as a framework for solving problems beyond sequence typing and gene detection. Chapter 4 presented the application of the STing algorithm in two different areas: (1) public health, and (2) environmental microbial genomics. In the first area, STing was used for virulence gene profiling of Shiga toxin-producing *Escherichia coli* (STEC) from WGS samples. Results showed that STing is more accurate than the PCR method, the gold-standard technique used by public health laboratories for characterizing virulence genes in STEC samples. STing had between 98% and 100% accuracy in characterizing the four genes used as markers for STEC determination (*stx1*, *stx2*, *eae*, and *ehxA*), compared to a PCR accuracy between 90% and 94%. Most importantly, and unlike the PCR technique, STing was able to detect novel genes in the analyzed STEC isolates. In the second area, the STing algorithm was expanded for *nifH* gene-based taxonomic classification of amplicon sequencing reads. The STing algorithm was implemented into a software utility for amplicon sequencing read classification – STing classifier – based on a *nifH* gene reference database. The STing classifier utility was able to correctly classify reads in nitrogen-fixing community samples simulated up to the lowest sequencing depth of 1x coverage. Importantly, results showed that full-length reference sequences of the gene *nifH* led to better taxonomy classification of the sequencing reads.

To summarize the main findings of my research, I developed an alignment-free algorithm based on *k*-mer frequencies that I further implemented into turn-key software applications for genome-enabled molecular epidemiology and environmental microbial genomics. The algorithm and its derived software applications showed to be more efficient, more accurate, and faster than the traditional alignment-based methods for NGS-based molecular epidemiology. Moreover, the algorithm has the potential for application to other areas such as environmental microbial genomics. Lastly, my research contributed significantly to overcome the challenges associated with the adoption of genome-enabled approaches to molecular epidemiology in public health.

While the alignment-free framework developed in this research contributes to the efficient use of NGS data in molecular epidemiology, there exist several aspects that can be further explored and developed in the algorithm and software applications.

In the sequence typing and gene detection analysis, a significance metric would help to improve the accuracy of predicted alleles/genes. A significance metric based on the expected and observed number of *k*-mers matches could be included for allele/gene calling to disambiguate between sequences with the same *k*-mer frequency without requiring the calculation of allele coverage and *k*-mer depth for this purpose.

A useful feature of STing is the indication of possible novel alleles/genes in an analyzed sample. Alleles or genes reported with coverage below 100% represent the closest sequences in the database to the actual sequences present in the sample analyzed. Such cases are probable novel alleles that are not characterized but included in the database index. A potential exciting feature for inclusion in the applications would be the ability to report the reads from that generated the *k*-mer matches of a probable novel allele. This feature would allow the user to reconstruct the probable novel allele from the reported sequence reads using a computationally inexpensive local assembly.

The excellent performance of the STing algorithm for virulence gene profiling of STEC is another opportunity for development. A dedicated software application that implements the alignment-free algorithm and the additional parameters (coverage and *k*-mer fraction) used for detecting the virulence genes in STEC samples will be valuable for public health laboratories dedicated to foodborne and waterborne disease control and surveillance.

Finally, the algorithm for gene-based read classification can be refined to improve classification performance. As mentioned in Chapter 4, the read assignment in cases in which two or more probable sequences from the index database have the same *k*-mer frequency, can be improved using two possible options: (1) assigning the read equally (*i.e.*, 1/*n*) to all the *n* probable sequences with even frequency, or (2) assigning the read count to the least common ancestor of the n probable sequences in the taxonomic hierarchy. Also, it would be valuable to explore the impact on read classification of increasing the number of *k*-mers used for the read filtering step. Considering that the algorithm is intended for amplicon sequence data analysis, it is highly likely that all the reads analyzed are useful since they arise from the amplified and sequenced region of interest. Thus, the read filtering step could be omitted, but at the expense of more time for the analysis. Further exploration of this approach is needed to determine if there is an improvement in read classification that justifies using more time for analysis.

PUBLICATIONS

Espitia, H., Chande, A. T., Smith, H., Jordan, I. K., & Rishishwar, L. (2017). A method of sequence typing with in silico aptamers from a next generation sequencing platform. Patent application US15/726,005. https://patents.google.com/patent/US20190108308A1.

Espitia-Navarro, H. F., Rishishwar, L., Mayer, L. W., Jordan, I. K. (2019). Bioinformatics. In B. a. S. Budowle, S. and Morse, S. (Ed.), Microbial Forensics (3rd ed.): Academic Press.

Espitia-Navarro, H. F., Chande, A. T., Nagar, S. D., Smith, H., Jordan, I. K., & Rishishwar, L. (2019). STing: accurate and ultrafast genomic profiling with exact sequence matches. Biorxiv. doi:10.1101/855478

APPENDIX A. SUPPLEMENTARY DATA FOR CHAPTER 2

A.1 Pseudocode for database indexing

Input:

Configuration file that defines relative paths to the required files:

1. Allele sequence file(s): Multi-fasta format file(s) in which the sequence description corresponds to the locus name and allele number for each locus in the MLST scheme.
2. Profiles file: Tab separated values file with the allelic profiles that define each ST.

Output:

A set of files that defines the database:

1. Allele ESA index files: `prefix.ali`, `prefix.dat`, `prefix.ids`, `prefix.sa`, `prefix.txt.[0-N]`.
2. Profile ESA index files: `prefix.prof_idx.ali`, `prefix.prof_idx.dat`, `prefix.prof_idx.ids`, `prefix.prof_idx.sa`, `prefix.prof_idx.txt.[0-N]`,

where `prefix` is the database prefix defined by the user.

General algorithm

1. Load the config file
2. Load the loci sequence files
   2.1. Store the allele sequences for each locus
   2.2. Store the allele sequence ids
3. Load the profiles file
   3.1. Store the profiles table
   3.2. Build a list of the string representation of each profile that defines an ST
4. Create and save to disk an ESA index of the list of profile string representations

5. Create and save to disk an ESA index of the allele sequences
6. Save auxiliary information
    6.1. Save a loci table file
    6.2. Save an allele-loci id pairs file
    6.3. Save a profiles file
    6.4. Save an allele sequences ids file

Input files description

The database indexing stage requires a configuration file that define the location of the files that make up a typing scheme: the allele sequence files and a profile definition file. Files of the typing scheme can be downloaded from PubMLST or can be created by the user.

Configuration file

A plain text file that contains two sections to specify the paths to the files of the typing scheme. An example of a configuration file is the following:

```
[loci]
# this is a comment line
abcZ     Neisseria_sp/abcZ.fa
adk      Neisseria_sp/adk.fa
aroE     Neisseria_sp/aroE.fa
fumC     Neisseria_sp/fumC.fa
gdh      Neisseria_sp/gdh.fa
pdhC     Neisseria_sp/pdhC.fa
pgm      Neisseria_sp/pgm.fa

[profile]
profile  Neisseria_sp/neisseria.txt
```

Each header (words between characters `[]`) define a section that contains paths to the input files. Each row of the section `[loci]` defines the allele sequence file for each locus of a typing scheme (name and file path). The section `[profile]` only has a row that specifies the profile definition file (name and file path). The configuration file must fulfill the following requirements:

- Sections headers are mandatory (`[loci]` and `[profile]`).
- Values defined in each row must be separated by TAB character.
- Blank lines and comments (lines starting with `#`) are ignored.
- File paths are relative to the location of the config file. In the example shown above, the folder `Neisseria_sp` is located at the same directory level of the configuration file.

Allele sequence file

Each allele sequence file (one for each locus in the typing scheme) is a multi-fasta file in which the description for each allele sequence is formed by the locus name with the allele number. An example of the *abcZ* allele sequence is the following:

```
>abcZ_1
TTTGATACTGTTGCCGTAC...
>abcZ_2
TTTGATACCGTTGCCGAAA...
>abcZ_3
TTTGATACCGTTGCGAACC...
>abcZ_4
TTTGATACCGTTGCCACGT...
```

Profile definition file

The profile definition file is a tab separated file that contains the ST and the allele profile corresponding to the ST. An example of the profile definition file is shown below:

```
ST	abcZ	adk	aroE	fumC	gdh	pdhC	pgm
1	1	3	1	1	1	1	3
2	1	3	4	7	1	1	3
3	1	3	1	1	1	23	13
4	1	3	3	1	4	2	3
```

A.2 Pseudocode for sequence typing

Input:

1. Raw FASTQ sequencing reads (single end or paired end)
2. Database index files

Output:

1. Allelic profile and associated sequence type
2. Total number of *k*-mer matches and reads processed
3. Optional information:
    a. Normalized counts of *k*-mer matches
    b. Coverage of each allele
    c. Mean *k*-mer depth
    d. Per-base *k*-mer depth file

Table 14. Whole genome sequencing samples used in the study for testing the sequence typing feature in STing.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
	Campylobacter jejuni – Genome size: **1,641,481 bp; Sequencing platform: Illumina**							
1	ERR1343082	125	894	2,866,336	218.3	1	0	0
2	SRR1723386	233	942	1,823,522	258.8	1	0	0
3	SRR1724281	199	1,260	2,778,714	337.6	1	0	0
4	SRR1769300	221	549	1,112,826	150.0	1	0	0
5	SRR1952324	100	845	2,535,868	154.5	1	0	0
6	SRR1952328	100	571	1,716,560	104.6	1	0	0
7	SRR1975125	100	1,137	3,409,294	207.7	1	0	0
8	SRR1981464	175	492	1,214,346	129.3	1	0	0
9	SRR1993476	100	2,681	8,067,062	491.5	1	0	0
10	SRR1993502	100	483	1,452,478	88.5	1	0	0
	Chlamydia trachomatis – Genome size: **1,038,842 bp; Sequencing platform: Illumina**							
11	ERR211059	100	881	3,355,454	323.0	1	0	0
12	ERR211062	100	1,035	3,936,668	378.9	1	0	0
13	ERR278147	100	1,342	5,095,590	490.5	1	0	0
14	ERR278160	100	1,039	3,950,456	380.3	1	0	0
15	ERR278188	100	1,386	5,260,678	506.4	1	0	0
16	ERR278190	100	441	1,692,518	162.9	1	0	0
17	ERR278214	100	1,105	4,198,314	404.1	1	0	0
18	ERR386224	150	1,059	2,960,372	427.5	1	0	0
19	ERR386229	150	1,002	2,804,346	404.9	1	0	0
20	ERR386231	150	881	2,466,280	356.1	1	0	0
	Neisseria meningitidis – Genome size: **2,272,360 bp; Sequencing platform: Illumina**							
21	ERR026496	54	1,096	7,184,442	178.0	0	1	0
22	ERR026500	54	1,250	8,192,192	202.9	0	1	0
23	ERR026503	54	340	2,233,096	55.3	0	1	0
24	ERR026504	54	304	1,991,980	49.3	0	1	0
25	ERR026505	54	392	2,573,244	63.7	0	1	0
26	ERR026507	54	412	2,709,874	67.1	0	1	0
27	ERR026509	54	762	4,998,372	123.8	0	1	0
28	ERR026510	54	522	3,408,742	84.4	0	1	0
29	ERR026511	54	282	1,837,012	45.5	0	1	0
30	ERR026512	54	444	2,896,590	71.8	0	1	0
31	ERR026514	54	498	3,268,300	81.0	0	1	0
32	ERR026516	54	400	2,623,604	65.0	0	1	0
33	ERR026518	54	472	3,095,454	76.7	0	1	0
34	ERR026519	54	270	1,774,294	42.2	0	1	0
35	ERR026524	54	350	2,281,138	56.5	0	1	0
36	ERR026526	54	878	5,759,248	142.7	0	1	0
37	ERR026529	54	172	1,122,980	27.8	0	1	0
38	ERR026533	54	448	2,945,852	73.0	0	1	0
39	ERR027242	54	918	5,974,310	148.0	0	1	0
40	ERR027244	54	932	6,028,512	149.3	0	1	0
41	ERR027245	54	178	1,148,086	28.4	0	1	0
42	ERR027247	54	952	6,195,372	153.5	0	1	0
43	ERR027248	54	960	6,257,436	155.0	0	1	0
44	ERR027249	54	1,102	7,173,706	177.7	0	1	0
45	ERR027250	54	36	224,638	5.6	0	1	0
46	ERR027251	54	994	6,479,664	160.5	0	1	0
47	ERR027252	54	1,476	9,611,612	238.1	0	1	0
48	ERR028371	54	932	6,032,308	149.4	0	1	0
49	ERR028376	54	1,162	7,567,326	187.4	0	1	0
50	ERR028378	54	326	2,128,172	52.7	0	1	0
51	ERR033097	54	596	3,860,334	95.6	0	1	0
52	ERR033098	54	540	3,500,508	86.7	0	1	0
53	ERR033099	54	1,010	6,535,968	161.9	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
54	ERR033101	54	670	4,361,792	108.0	0	1	0
55	ERR033102	54	588	3,827,788	94.8	0	1	0
56	ERR033103	54	426	2,777,428	68.8	0	1	0
57	ERR033105	54	418	2,728,966	67.6	0	1	0
58	ERR033107	54	756	4,926,230	122.0	0	1	0
59	ERR036060	76	1,048	5,356,738	186.7	0	1	0
60	ERR036061	76	1,330	6,771,442	236.1	0	1	0
61	ERR036062	76	920	4,689,224	163.5	0	1	0
62	ERR036063	76	1,650	8,401,526	292.9	0	1	0
63	ERR036064	76	1,392	7,122,210	248.3	0	1	0
64	ERR036065	76	1,266	6,471,986	225.6	0	1	0
65	ERR036066	76	1,326	6,785,994	236.6	0	1	0
66	ERR036067	76	1,078	5,519,278	192.4	0	1	0
67	ERR036068	76	1,654	8,461,934	295.0	0	1	0
68	ERR036069	76	1,576	8,064,462	281.1	0	1	0
69	ERR036070	76	1,268	6,485,142	226.1	0	1	0
70	ERR036071	76	1,222	6,247,392	217.8	0	1	0
71	ERR036073	76	1,178	6,025,722	210.1	0	1	0
72	ERR036074	76	1,126	5,727,406	199.7	0	1	0
73	ERR036075	76	1,038	5,281,942	184.1	0	1	0
74	ERR036076	76	1,408	7,168,474	249.9	0	1	0
75	ERR036077	76	1,440	7,361,986	256.7	0	1	0
76	ERR036078	76	2,624	13,420,062	467.9	0	1	0
77	ERR036079	76	1,596	8,165,946	284.7	0	1	0
78	ERR036080	76	1,136	5,809,028	202.5	0	1	0
79	ERR036081	76	1,174	6,009,196	209.5	0	1	0
80	ERR036082	76	1,428	7,307,602	254.8	0	1	0
81	ERR036083	76	1,458	7,456,970	260.0	0	1	0
82	ERR036084	76	1,174	6,004,598	209.3	0	1	0
83	ERR036086	76	1,096	5,605,668	195.4	0	1	0
84	ERR036090	76	1,622	8,296,582	289.2	0	1	0
85	ERR036091	76	2,102	10,746,520	374.6	0	1	0
86	ERR036093	76	1,228	6,277,710	218.9	0	1	0
87	ERR036094	76	1,070	5,477,104	190.9	0	1	0
88	ERR036099	76	292	1,495,004	52.1	0	1	0
89	ERR036100	76	3,756	19,101,682	665.9	0	1	0
90	ERR036101	76	1,054	5,367,066	187.1	0	1	0
91	ERR036102	76	680	3,466,366	120.8	0	1	0
92	ERR036103	76	338	1,732,120	60.4	0	1	0
93	ERR036104	76	134	679,262	23.7	0	1	0
94	ERR036105	76	242	1,241,630	43.3	0	1	0
95	ERR036106	76	600	3,075,856	107.2	0	1	0
96	ERR036107	76	496	2,540,162	88.6	0	1	0
97	ERR036108	76	754	3,860,916	134.6	0	1	0
98	ERR036109	76	798	4,083,756	142.4	0	1	0
99	ERR036110	76	1,572	8,043,458	280.4	0	1	0
100	ERR036112	76	804	4,109,146	143.3	0	1	0
101	ERR036113	76	376	1,922,176	67.0	0	1	0
102	ERR036114	76	920	4,688,750	163.5	0	1	0
103	ERR036115	76	1,014	5,163,356	180.0	0	1	0
104	ERR036116	76	674	3,445,936	120.1	0	1	0
105	ERR036118	76	1,248	6,385,840	222.6	0	1	0
106	ERR036119	76	924	4,730,598	164.9	0	1	0
107	ERR036120	76	280	1,437,564	50.1	0	1	0
108	ERR036121	76	2,070	10,583,302	369.0	0	1	0
109	ERR036122	76	1,016	5,195,548	181.1	0	1	0
110	ERR036123	76	1,058	5,413,462	188.7	0	1	0
111	ERR063493	75	1,998	10,219,774	351.6	0	1	0
112	ERR063494	75	1,680	8,597,008	295.8	0	1	0
113	ERR063495	75	1,664	8,516,726	293.0	0	1	0
114	ERR063496	75	2,134	10,867,738	373.9	0	1	0
115	ERR063497	75	1,804	9,186,160	316.0	0	1	0
116	ERR063498	75	2,126	10,820,598	372.3	0	1	0
117	ERR063500	75	2,274	11,576,502	398.3	0	1	0
118	ERR063501	75	1,864	9,496,046	326.7	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
119	ERR063502	75	1,914	9,751,066	335.5	0	1	0
120	ERR063503	75	2,238	11,393,562	392.0	0	1	0
121	ERR086223	75	3,370	16,991,290	584.6	0	1	0
122	ERR086224	75	7,404	37,243,524	1,281.3	0	1	0
123	ERR086225	75	3,970	20,015,292	688.6	0	1	0
124	ERR086226	75	5,986	30,134,268	1,036.7	0	1	0
125	ERR086227	75	4,150	20,919,010	719.7	0	1	0
126	ERR086228	75	3,784	19,077,050	656.3	0	1	0
127	ERR086229	75	5,154	25,951,118	892.8	0	1	0
128	ERR086230	75	2,968	14,966,392	514.9	0	1	0
129	ERR086231	75	6,158	30,988,916	1,066.1	0	1	0
130	ERR086232	75	4,660	23,473,232	807.6	0	1	0
131	ERR086233	75	3,022	15,234,254	524.1	0	1	0
132	ERR086234	75	4,416	22,252,204	765.6	0	1	0
133	ERR133682	100	870	3,563,966	163.5	0	1	0
134	ERR133683	100	1,018	4,171,738	191.4	0	1	0
135	ERR133684	100	1,154	4,730,932	217.0	0	1	0
136	ERR133685	100	1,388	5,685,566	260.8	0	1	0
137	ERR133686	100	1,078	4,417,878	202.7	0	1	0
138	ERR133687	100	944	3,870,846	177.6	0	1	0
139	ERR133688	100	1,046	4,290,854	196.8	0	1	0
140	ERR133689	100	1,098	4,502,914	206.6	0	1	0
141	ERR133690	100	1,020	4,178,958	191.7	0	1	0
142	ERR133691	100	914	3,731,952	171.2	0	1	0
143	ERR133692	100	1,010	4,122,212	189.1	0	1	0
144	ERR133693	100	868	3,548,300	162.8	0	1	0
145	ERR133694	100	1,130	4,617,132	211.8	0	1	0
146	ERR133695	100	954	3,892,418	178.6	0	1	0
147	ERR133696	100	900	3,674,296	168.5	0	1	0
148	ERR133697	100	1,332	5,439,540	249.5	0	1	0
149	ERR133698	100	926	3,782,582	173.5	0	1	0
150	ERR133699	100	884	3,609,870	165.6	0	1	0
151	ERR133700	100	918	3,745,896	171.8	0	1	0
152	ERR133701	100	730	2,981,794	136.8	0	1	0
153	ERR133702	100	770	3,145,396	144.3	0	1	0
154	ERR133703	100	824	3,362,400	154.2	0	1	0
155	ERR133704	100	982	4,014,266	184.1	0	1	0
156	ERR133705	100	860	3,516,726	161.3	0	1	0
157	ERR133706	100	1,052	4,292,428	196.9	0	1	0
158	ERR133707	100	952	3,885,336	178.2	0	1	0
159	ERR133708	100	722	2,949,800	135.3	0	1	0
160	ERR133709	100	862	3,520,396	161.5	0	1	0
161	ERR133710	100	946	3,862,046	177.2	0	1	0
162	ERR133711	100	758	3,100,486	142.2	0	1	0
163	ERR133712	100	848	3,465,726	159.0	0	1	0
164	ERR133713	100	802	3,273,688	150.2	0	1	0
165	ERR133714	100	762	3,110,760	142.7	0	1	0
166	ERR133715	100	800	3,272,202	150.1	0	1	0
167	ERR133716	100	800	3,266,834	149.9	0	1	0
168	ERR133717	100	736	3,008,710	138.0	0	1	0
169	ERR133718	100	1,008	4,113,542	188.7	0	1	0
170	ERR133719	100	1,018	4,154,794	190.6	0	1	0
171	ERR133720	100	838	3,423,444	157.0	0	1	0
172	ERR133721	100	904	3,693,024	169.4	0	1	0
173	ERR133722	100	1,032	4,213,714	193.3	0	1	0
174	ERR133723	100	922	3,767,750	172.8	0	1	0
175	ERR133724	100	894	3,648,528	167.4	0	1	0
176	ERR133725	100	852	3,484,308	159.8	0	1	0
177	ERR133726	100	744	3,037,272	139.3	0	1	0
178	ERR133727	100	1,040	4,249,256	194.9	0	1	0
179	ERR133728	100	696	2,841,130	130.3	0	1	0
180	ERR133729	100	872	3,561,390	163.4	0	1	0
181	ERR133730	100	1,224	4,995,770	229.2	0	1	0
182	ERR133731	100	1,020	4,168,936	191.2	0	1	0
183	ERR133732	100	1,096	4,475,098	205.3	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
184	ERR133733	100	912	3,721,224	170.7	0	1	0
185	ERR133734	100	896	3,660,466	167.9	0	1	0
186	ERR133735	100	862	3,520,736	161.5	0	1	0
187	ERR133736	100	834	3,405,354	156.2	0	1	0
188	ERR133737	100	740	3,023,658	138.7	0	1	0
189	ERR133738	100	816	3,336,520	153.1	0	1	0
190	ERR133739	100	718	2,934,034	134.6	0	1	0
191	ERR133740	100	826	3,377,872	154.9	0	1	0
192	ERR133741	100	752	3,074,774	141.0	0	1	0
193	ERR133742	100	996	4,071,052	186.7	0	1	0
194	ERR133743	100	1,216	4,968,084	227.9	0	1	0
195	ERR133744	100	918	3,750,466	172.0	0	1	0
196	ERR133745	100	1,016	4,148,412	190.3	0	1	0
197	ERR133746	100	896	3,659,350	167.9	0	1	0
198	ERR133747	100	852	3,484,032	159.8	0	1	0
199	ERR133748	100	846	3,457,290	158.6	0	1	0
200	ERR133749	100	818	3,339,864	153.2	0	1	0
201	ERR133750	100	742	3,029,468	139.0	0	1	0
202	ERR133751	100	768	3,137,096	143.9	0	1	0
203	ERR133752	100	908	3,711,054	170.2	0	1	0
204	ERR133753	100	1,008	4,113,506	188.7	0	1	0
205	ERR133754	100	752	3,071,332	140.9	0	1	0
206	ERR133755	100	968	3,954,050	181.4	0	1	0
207	ERR133756	100	848	3,467,788	159.1	0	1	0
208	ERR133757	100	1,078	4,399,362	201.8	0	1	0
209	ERR133758	100	758	3,099,220	142.2	0	1	0
210	ERR133759	100	664	2,716,300	124.6	0	1	0
211	ERR133760	100	716	2,928,444	134.3	0	1	0
212	ERR133761	100	882	3,601,928	165.2	0	1	0
213	ERR133762	100	772	3,151,886	144.6	0	1	0
214	ERR133763	100	776	3,168,482	145.3	0	1	0
215	ERR133764	100	854	3,489,490	160.1	0	1	0
216	ERR133765	100	720	2,940,090	134.9	0	1	0
217	ERR133766	100	680	2,778,670	127.5	0	1	0
218	ERR133767	100	742	3,034,446	139.2	0	1	0
219	ERR133768	100	856	3,500,566	160.6	0	1	0
220	ERR133769	100	770	3,146,000	144.3	0	1	0
221	ERR133770	100	766	3,132,248	143.7	0	1	0
222	ERR133771	100	1,014	4,141,128	190.0	0	1	0
223	ERR133772	100	682	2,786,270	127.8	0	1	0
224	ERR133773	100	828	3,379,256	155.0	0	1	0
225	ERR133774	100	708	2,896,094	132.8	0	1	0
226	ERR133775	100	806	3,294,410	151.1	0	1	0
227	ERR133776	100	654	2,671,938	122.6	0	1	0
228	ERR133777	100	672	2,745,302	125.9	0	1	0
229	ERR133778	100	4,942	20,163,292	924.9	0	1	0
230	ERR133779	100	5,046	20,588,544	944.4	0	1	0
231	ERR133780	100	7,072	28,822,954	1,322.2	0	1	0
232	ERR133781	100	5,998	24,457,486	1,121.9	0	1	0
233	ERR133782	100	4,824	19,683,856	902.9	0	1	0
234	ERR133783	100	4,228	17,250,638	791.3	0	1	0
235	ERR133784	100	5,430	22,152,760	1,016.2	0	1	0
236	ERR133785	100	5,494	22,405,994	1,027.8	0	1	0
237	ERR133786	100	6,212	25,331,016	1,162.0	0	1	0
238	ERR133787	100	5,586	22,694,010	1,041.0	0	1	0
239	ERR133788	100	6,198	25,173,852	1,154.8	0	1	0
240	ERR133789	100	6,266	25,444,550	1,167.2	0	1	0
241	ERR133790	100	5,274	21,433,642	983.2	0	1	0
242	ERR133791	100	4,676	19,010,466	872.0	0	1	0
243	ERR133792	100	5,222	21,225,052	973.6	0	1	0
244	ERR133793	100	4,778	19,424,464	891.0	0	1	0
245	ERR133794	100	4,902	19,925,702	914.0	0	1	0
246	ERR137095	100	1,136	4,638,038	212.8	0	1	0
247	ERR137096	100	1,430	5,834,448	267.6	0	1	0
248	ERR137097	100	1,510	6,167,706	282.9	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
249	ERR137098	100	1,516	6,187,894	283.8	0	1	0
250	ERR137099	100	1,434	5,858,340	268.7	0	1	0
251	ERR137100	100	1,204	4,914,272	225.4	0	1	0
252	ERR137101	100	1,356	5,538,380	254.1	0	1	0
253	ERR137102	100	1,488	6,071,300	278.5	0	1	0
254	ERR137103	100	1,370	5,597,526	256.8	0	1	0
255	ERR137104	100	1,230	5,006,030	229.6	0	1	0
256	ERR137105	100	1,368	5,562,932	255.2	0	1	0
257	ERR137106	100	1,438	5,845,698	268.2	0	1	0
258	ERR137107	100	1,342	5,456,798	250.3	0	1	0
259	ERR137108	100	1,076	4,380,774	201.0	0	1	0
260	ERR137109	100	1,346	5,470,454	250.9	0	1	0
261	ERR137110	100	1,350	5,493,748	252.0	0	1	0
262	ERR137111	100	1,370	5,569,124	255.5	0	1	0
263	ERR137112	100	1,068	4,346,644	199.4	0	1	0
264	ERR137113	100	1,194	4,853,038	222.6	0	1	0
265	ERR137114	100	900	3,658,534	167.8	0	1	0
266	ERR137115	100	1,046	4,257,408	195.3	0	1	0
267	ERR137116	100	996	4,051,002	185.8	0	1	0
268	ERR137117	100	1,276	5,188,460	238.0	0	1	0
269	ERR137118	100	1,348	5,480,184	251.4	0	1	0
270	ERR137119	100	1,172	4,766,836	218.7	0	1	0
271	ERR137120	100	1,342	5,455,944	250.3	0	1	0
272	ERR137121	100	918	3,733,002	171.2	0	1	0
273	ERR137122	100	944	3,837,426	176.0	0	1	0
274	ERR137123	100	882	3,584,856	164.4	0	1	0
275	ERR137124	100	1,230	5,003,702	229.5	0	1	0
276	ERR137125	100	972	3,956,220	181.5	0	1	0
277	ERR137126	100	1,044	4,244,308	194.7	0	1	0
278	ERR137127	100	872	3,548,278	162.8	0	1	0
279	ERR137128	100	998	4,058,004	186.1	0	1	0
280	ERR137129	100	942	3,835,630	175.9	0	1	0
281	ERR137130	100	1,022	4,158,492	190.8	0	1	0
282	ERR137131	100	1,172	4,766,810	218.7	0	1	0
283	ERR137132	100	1,086	4,416,298	202.6	0	1	0
284	ERR137133	100	1,024	4,166,944	191.1	0	1	0
285	ERR137134	100	1,050	4,271,304	195.9	0	1	0
286	ERR137135	100	1,082	4,397,606	201.7	0	1	0
287	ERR137136	100	1,254	5,101,324	234.0	0	1	0
288	ERR137137	100	1,360	5,532,482	253.8	0	1	0
289	ERR137138	100	1,262	5,130,246	235.3	0	1	0
290	ERR137139	100	1,330	5,410,390	248.2	0	1	0
291	ERR137140	100	1,138	4,630,412	212.4	0	1	0
292	ERR137141	100	1,032	4,195,272	192.4	0	1	0
293	ERR137142	100	1,324	5,386,182	247.1	0	1	0
294	ERR137143	100	1,510	6,140,390	281.7	0	1	0
295	ERR137144	100	1,284	5,221,324	239.5	0	1	0
296	ERR137145	100	1,052	4,276,826	196.2	0	1	0
297	ERR137146	100	1,242	5,048,296	231.6	0	1	0
298	ERR137147	100	1,156	4,701,692	215.7	0	1	0
299	ERR137148	100	946	3,846,542	176.4	0	1	0
300	ERR137149	100	1,314	5,342,466	245.1	0	1	0
301	ERR137150	100	1,238	5,038,764	231.1	0	1	0
302	ERR137151	100	1,080	4,392,492	201.5	0	1	0
303	ERR137152	100	1,468	5,971,844	273.9	0	1	0
304	ERR137153	100	1,116	4,542,548	208.4	0	1	0
305	ERR137154	100	1,326	5,391,684	247.3	0	1	0
306	ERR137155	100	758	3,088,624	141.7	0	1	0
307	ERR137156	100	1,284	5,225,632	239.7	0	1	0
308	ERR137157	100	936	3,810,180	174.8	0	1	0
309	ERR137158	100	1,076	4,379,018	200.9	0	1	0
310	ERR137159	100	1,120	4,559,828	209.2	0	1	0
311	ERR137160	100	1,080	4,394,496	201.6	0	1	0
312	ERR137161	100	976	3,974,762	182.3	0	1	0
313	ERR137162	100	1,470	5,977,146	274.2	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
314	ERR137163	100	1,116	4,542,766	208.4	0	1	0
315	ERR137164	100	1,122	4,567,312	209.5	0	1	0
316	ERR137165	100	1,054	4,284,096	196.5	0	1	0
317	ERR137166	100	1,074	4,368,986	200.4	0	1	0
318	ERR137167	100	1,138	4,630,640	212.4	0	1	0
319	ERR137168	100	1,174	4,771,774	218.9	0	1	0
320	ERR137169	100	832	3,389,258	155.5	0	1	0
321	ERR137170	100	1,228	4,990,798	228.9	0	1	0
322	ERR137171	100	1,042	4,235,824	194.3	0	1	0
323	ERR137172	100	900	3,664,132	168.1	0	1	0
324	ERR137173	100	1,032	4,198,236	192.6	0	1	0
325	ERR137174	100	998	4,063,636	186.4	0	1	0
326	ERR137175	100	1,072	4,357,178	199.9	0	1	0
327	ERR137176	100	1,228	4,990,984	228.9	0	1	0
328	ERR137177	100	1,006	4,092,994	187.8	0	1	0
329	ERR137178	100	1,058	4,305,464	197.5	0	1	0
330	ERR137179	100	1,284	5,219,880	239.4	0	1	0
331	ERR137180	100	1,390	5,654,378	259.4	0	1	0
332	ERR137181	100	1,052	4,281,308	196.4	0	1	0
333	ERR137182	100	1,128	4,587,514	210.4	0	1	0
334	ERR137183	100	1,124	4,575,024	209.9	0	1	0
335	ERR137184	100	1,284	5,220,970	239.5	0	1	0
336	ERR137185	100	1,244	5,060,566	232.1	0	1	0
337	ERR137186	100	976	3,967,938	182.0	0	1	0
338	ERR137187	100	1,090	4,435,592	203.5	0	1	0
339	ERR137188	100	1,028	4,182,526	191.9	0	1	0
340	ERR137189	100	1,104	4,490,958	206.0	0	1	0
341	ERR137190	100	1,098	4,466,536	204.9	0	1	0
342	ERR160731	100	1,958	7,995,046	366.7	0	1	0
343	ERR160751	100	1,760	7,158,888	328.4	0	1	0
344	ERR160754	100	1,778	7,231,922	331.7	0	1	0
345	ERR160759	100	2,188	8,896,048	408.1	0	1	0
346	ERR160773	100	1,770	7,197,716	330.2	0	1	0
347	ERR160786	100	1,888	7,674,640	352.0	0	1	0
348	ERR170738	100	556	2,283,394	104.7	0	1	0
349	ERR170739	100	500	2,057,468	94.4	0	1	0
350	ERR170740	100	642	2,642,304	121.2	0	1	0
351	ERR170741	100	558	2,298,372	105.4	0	1	0
352	ERR170742	100	556	2,289,850	105.0	0	1	0
353	ERR170743	100	654	2,687,210	123.3	0	1	0
354	ERR170744	100	460	1,893,272	86.8	0	1	0
355	ERR170745	100	508	2,091,698	95.9	0	1	0
356	ERR170746	100	484	1,990,650	91.3	0	1	0
357	ERR170747	100	452	1,852,726	85.0	0	1	0
358	ERR170748	100	600	2,454,148	112.6	0	1	0
359	ERR170749	100	490	2,003,760	91.9	0	1	0
360	ERR170750	100	522	2,135,396	98.0	0	1	0
361	ERR170751	100	572	2,339,610	107.3	0	1	0
362	ERR170752	100	622	2,544,810	116.7	0	1	0
363	ERR170753	100	572	2,341,180	107.4	0	1	0
364	ERR170754	100	510	2,093,236	96.0	0	1	0
365	ERR170755	100	688	2,818,670	129.3	0	1	0
366	ERR170756	100	440	1,801,468	82.6	0	1	0
367	ERR170757	100	362	1,481,568	68.0	0	1	0
368	ERR170758	100	558	2,284,146	104.8	0	1	0
369	ERR170759	100	556	2,278,036	104.5	0	1	0
370	ERR170760	100	642	2,629,994	120.6	0	1	0
371	ERR170761	100	594	2,432,032	111.6	0	1	0
372	ERR170762	100	522	2,136,428	98.0	0	1	0
373	ERR170763	100	722	2,956,440	135.6	0	1	0
374	ERR170764	100	526	2,157,844	99.0	0	1	0
375	ERR170765	100	402	1,644,608	75.4	0	1	0
376	ERR170766	100	474	1,942,202	89.1	0	1	0
377	ERR170767	100	582	2,379,956	109.2	0	1	0
378	ERR170768	100	456	1,867,132	85.6	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
379	ERR170769	100	458	1,877,342	86.1	0	1	0
380	ERR170770	100	566	2,322,302	106.5	0	1	0
381	ERR170771	100	512	2,099,784	96.3	0	1	0
382	ERR170772	100	574	2,353,918	108.0	0	1	0
383	ERR170773	100	682	2,790,322	128.0	0	1	0
384	ERR170774	100	516	2,110,356	96.8	0	1	0
385	ERR170775	100	546	2,238,472	102.7	0	1	0
386	ERR170776	100	470	1,929,148	88.5	0	1	0
387	ERR170777	100	570	2,337,390	107.2	0	1	0
388	ERR170778	100	400	1,635,006	75.0	0	1	0
389	ERR170779	100	568	2,329,632	106.9	0	1	0
390	ERR170780	100	444	1,820,324	83.5	0	1	0
391	ERR170781	100	840	3,435,924	157.6	0	1	0
392	ERR170782	100	580	2,373,782	108.9	0	1	0
393	ERR170783	100	416	1,700,840	78.0	0	1	0
394	ERR170784	100	544	2,231,868	102.4	0	1	0
395	ERR170785	100	568	2,323,438	106.6	0	1	0
396	ERR170786	100	566	2,322,388	106.5	0	1	0
397	ERR170787	100	474	1,943,586	89.2	0	1	0
398	ERR170788	100	676	2,770,280	127.1	0	1	0
399	ERR170789	100	562	2,299,520	105.5	0	1	0
400	ERR170790	100	416	1,701,746	78.1	0	1	0
401	ERR170791	100	602	2,468,748	113.2	0	1	0
402	ERR170792	100	508	2,084,124	95.6	0	1	0
403	ERR170793	100	580	2,373,316	108.9	0	1	0
404	ERR170794	100	680	2,785,140	127.8	0	1	0
405	ERR170795	100	532	2,179,432	100.0	0	1	0
406	ERR170796	100	618	2,528,872	116.0	0	1	0
407	ERR170797	100	700	2,865,964	131.5	0	1	0
408	ERR170798	100	352	1,444,280	66.3	0	1	0
409	ERR170799	100	586	2,397,146	110.0	0	1	0
410	ERR170800	100	522	2,138,852	98.1	0	1	0
411	ERR170801	100	598	2,452,332	112.5	0	1	0
412	ERR170802	100	502	2,055,706	94.3	0	1	0
413	ERR170803	100	410	1,677,740	77.0	0	1	0
414	ERR170804	100	510	2,090,856	95.9	0	1	0
415	ERR170805	100	620	2,539,768	116.5	0	1	0
416	ERR170806	100	458	1,872,550	85.9	0	1	0
417	ERR170807	100	532	2,178,168	99.9	0	1	0
418	ERR170808	100	576	2,358,244	108.2	0	1	0
419	ERR170809	100	622	2,550,138	117.0	0	1	0
420	ERR170810	100	646	2,642,044	121.2	0	1	0
421	ERR170811	100	466	1,912,838	87.7	0	1	0
422	ERR170812	100	676	2,764,940	126.8	0	1	0
423	ERR170813	100	546	2,232,844	102.4	0	1	0
424	ERR170814	100	722	2,953,618	135.5	0	1	0
425	ERR170815	100	520	2,127,288	97.6	0	1	0
426	ERR170816	100	520	2,127,948	97.6	0	1	0
427	ERR170817	100	548	2,244,352	103.0	0	1	0
428	ERR170818	100	520	2,127,748	97.6	0	1	0
429	ERR170819	100	574	2,351,956	107.9	0	1	0
430	ERR170820	100	614	2,513,228	115.3	0	1	0
431	ERR170821	100	658	2,694,938	123.6	0	1	0
432	ERR170822	100	408	1,670,856	76.6	0	1	0
433	ERR170823	100	662	2,708,188	124.2	0	1	0
434	ERR170824	100	650	2,663,536	122.2	0	1	0
435	ERR170825	100	514	2,108,538	96.7	0	1	0
436	ERR170826	100	498	2,038,992	93.5	0	1	0
437	ERR170827	100	442	1,813,282	83.2	0	1	0
438	ERR170828	100	598	2,453,322	112.5	0	1	0
439	ERR170829	100	560	2,295,864	105.3	0	1	0
440	ERR170830	100	466	1,907,136	87.5	0	1	0
441	ERR170831	100	530	2,168,698	99.5	0	1	0
442	ERR170832	100	550	2,250,680	103.2	0	1	0
443	ERR170833	100	664	2,719,168	124.7	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
444	ERR170834	100	444	1,827,282	83.8	0	1	0
445	ERR170835	100	472	1,938,090	88.9	0	1	0
446	ERR170836	100	586	2,409,172	110.5	0	1	0
447	ERR170837	100	680	2,795,528	128.2	0	1	0
448	ERR170838	100	830	3,408,180	156.3	0	1	0
449	ERR170839	100	666	2,734,118	125.4	0	1	0
450	ERR170840	100	722	2,970,356	136.3	0	1	0
451	ERR170841	100	574	2,361,520	108.3	0	1	0
452	ERR170842	100	774	3,184,986	146.1	0	1	0
453	ERR170843	100	834	3,414,318	156.6	0	1	0
454	ERR170844	100	1,470	6,010,960	275.7	0	1	0
455	ERR170845	100	548	2,248,424	103.1	0	1	0
456	ERR170846	100	586	2,402,586	110.2	0	1	0
457	ERR170847	100	352	1,439,964	66.1	0	1	0
458	ERR170848	100	552	2,257,792	103.6	0	1	0
459	ERR170849	100	478	1,960,154	89.9	0	1	0
460	ERR170850	100	442	1,810,334	83.0	0	1	0
461	ERR170851	100	314	1,283,196	58.9	0	1	0
462	ERR170852	100	320	1,311,882	60.2	0	1	0
463	ERR170853	100	512	2,099,902	96.3	0	1	0
464	ERR170854	100	356	1,457,374	66.9	0	1	0
465	ERR170855	100	460	1,880,838	86.3	0	1	0
466	ERR170856	100	728	2,983,460	136.9	0	1	0
467	ERR170857	100	560	2,291,126	105.1	0	1	0
468	ERR170858	100	504	2,061,550	94.6	0	1	0
469	ERR170859	100	380	1,556,988	71.4	0	1	0
470	ERR170860	100	280	1,143,376	52.4	0	1	0
471	ERR170861	100	654	2,681,036	123.0	0	1	0
472	ERR170862	100	458	1,877,814	86.1	0	1	0
473	ERR170863	100	500	2,050,618	94.1	0	1	0
474	ERR170864	100	390	1,599,574	73.4	0	1	0
475	ERR170865	100	366	1,496,200	68.6	0	1	0
476	ERR170866	100	420	1,719,648	78.9	0	1	0
477	ERR170867	100	444	1,821,648	83.6	0	1	0
478	ERR170868	100	488	2,001,326	91.8	0	1	0
479	ERR170869	100	404	1,657,868	76.0	0	1	0
480	ERR170871	100	366	1,499,028	68.8	0	1	0
481	ERR170872	100	342	1,396,690	64.1	0	1	0
482	ERR170873	100	374	1,534,306	70.4	0	1	0
483	ERR170874	100	352	1,441,482	66.1	0	1	0
484	ERR170875	100	394	1,610,054	73.9	0	1	0
485	ERR170876	100	316	1,291,660	59.3	0	1	0
486	ERR170877	100	294	1,204,662	55.3	0	1	0
487	ERR170878	100	390	1,597,464	73.3	0	1	0
488	ERR170879	100	428	1,755,134	80.5	0	1	0
489	ERR170880	100	456	1,868,974	85.7	0	1	0
490	ERR170881	100	452	1,851,126	84.9	0	1	0
491	ERR170882	100	464	1,897,162	87.0	0	1	0
492	ERR170883	100	382	1,565,546	71.8	0	1	0
493	ERR170884	100	362	1,481,592	68.0	0	1	0
494	ERR170885	100	412	1,688,210	77.4	0	1	0
495	ERR170886	100	486	1,992,614	91.4	0	1	0
496	ERR170887	100	400	1,639,848	75.2	0	1	0
497	ERR170888	100	344	1,408,480	64.6	0	1	0
498	ERR170889	100	330	1,350,770	62.0	0	1	0
499	ERR170890	100	358	1,463,818	67.1	0	1	0
500	ERR170891	100	406	1,666,278	76.4	0	1	0
501	ERR170892	100	386	1,584,320	72.7	0	1	0
502	ERR170893	100	534	2,186,442	100.3	0	1	0
503	ERR170894	100	418	1,712,304	78.5	0	1	0
504	ERR170895	100	324	1,327,014	60.9	0	1	0
505	ERR170896	100	392	1,605,882	73.7	0	1	0
506	ERR170897	100	540	2,210,844	101.4	0	1	0
507	ERR170898	100	334	1,365,290	62.6	0	1	0
508	ERR170899	100	386	1,578,104	72.4	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
509	ERR170900	100	570	2,338,508	107.3	0	1	0
510	ERR170901	100	386	1,578,670	72.4	0	1	0
511	ERR170902	100	370	1,516,336	69.6	0	1	0
512	ERR170903	100	348	1,421,678	65.2	0	1	0
513	ERR170904	100	498	2,036,576	93.4	0	1	0
514	ERR170905	100	440	1,798,966	82.5	0	1	0
515	ERR170906	100	458	1,876,052	86.1	0	1	0
516	ERR170907	100	300	1,229,668	56.4	0	1	0
517	ERR170908	100	362	1,478,840	67.8	0	1	0
518	ERR170909	100	670	2,744,080	125.9	0	1	0
519	ERR170910	100	384	1,573,566	72.2	0	1	0
520	ERR170911	100	342	1,404,426	64.4	0	1	0
521	ERR170912	100	346	1,420,286	65.2	0	1	0
522	ERR170913	100	388	1,587,310	72.8	0	1	0
523	ERR170914	100	348	1,428,824	65.5	0	1	0
524	ERR170915	100	402	1,642,952	75.4	0	1	0
525	ERR170916	100	500	2,047,420	93.9	0	1	0
526	ERR170917	100	406	1,665,810	76.4	0	1	0
527	ERR170918	100	344	1,408,140	64.6	0	1	0
528	ERR170919	100	468	1,916,988	87.9	0	1	0
529	ERR170920	100	472	1,931,084	88.6	0	1	0
530	ERR170921	100	500	2,048,712	94.0	0	1	0
531	ERR170922	100	346	1,420,744	65.2	0	1	0
532	ERR170923	100	436	1,783,682	81.8	0	1	0
533	ERR170924	100	566	2,317,308	106.3	0	1	0
534	ERR170925	100	396	1,623,926	74.5	0	1	0
535	ERR170926	100	330	1,349,168	61.9	0	1	0
536	ERR170927	100	468	1,921,600	88.1	0	1	0
537	ERR170928	100	530	2,173,470	99.7	0	1	0
538	ERR170929	100	508	2,082,126	95.5	0	1	0
539	ERR170930	100	530	2,178,666	99.9	0	1	0
540	ERR170931	100	530	2,177,988	99.9	0	1	0
541	ERR170932	100	542	2,226,138	102.1	0	1	0
542	ERR170933	100	564	2,316,554	106.3	0	1	0
543	ERR170934	100	586	2,408,470	110.5	0	1	0
544	ERR170935	100	572	2,353,072	107.9	0	1	0
545	ERR170936	100	538	2,211,264	101.4	0	1	0
546	ERR170937	100	534	2,193,886	100.6	0	1	0
547	ERR170938	100	524	2,155,446	98.9	0	1	0
548	ERR170939	100	554	2,268,656	104.1	0	1	0
549	ERR170940	100	598	2,449,226	112.3	0	1	0
550	ERR170941	100	586	2,402,398	110.2	0	1	0
551	ERR170942	100	570	2,335,596	107.1	0	1	0
552	ERR170943	100	590	2,414,542	110.8	0	1	0
553	ERR170944	100	646	2,646,022	121.4	0	1	0
554	ERR170945	100	650	2,664,738	122.2	0	1	0
555	ERR170946	100	590	2,415,556	110.8	0	1	0
556	ERR170947	100	578	2,370,614	108.7	0	1	0
557	ERR170948	100	608	2,487,378	114.1	0	1	0
558	ERR170949	100	502	2,056,884	94.4	0	1	0
559	ERR170950	100	582	2,386,970	109.5	0	1	0
560	ERR170951	100	622	2,548,614	116.9	0	1	0
561	ERR170952	100	584	2,390,812	109.7	0	1	0
562	ERR170953	100	544	2,228,820	102.2	0	1	0
563	ERR170954	100	554	2,268,938	104.1	0	1	0
564	ERR170955	100	550	2,256,520	103.5	0	1	0
565	ERR170956	100	570	2,334,120	107.1	0	1	0
566	ERR170957	100	636	2,606,670	119.6	0	1	0
567	ERR170958	100	612	2,506,734	115.0	0	1	0
568	ERR170959	100	572	2,341,724	107.4	0	1	0
569	ERR170960	100	592	2,428,686	111.4	0	1	0
570	ERR170961	100	576	2,362,552	108.4	0	1	0
571	ERR170962	100	504	2,063,600	94.7	0	1	0
572	ERR170963	100	600	2,460,126	112.8	0	1	0
573	ERR170964	100	626	2,562,380	117.5	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
574	ERR170965	100	520	2,133,296	97.9	0	1	0
575	ERR170966	100	550	2,255,898	103.5	0	1	0
576	ERR170967	100	530	2,174,024	99.7	0	1	0
577	ERR170968	100	596	2,439,112	111.9	0	1	0
578	ERR170969	100	590	2,413,494	110.7	0	1	0
579	ERR170970	100	596	2,442,018	112.0	0	1	0
580	ERR170971	100	576	2,357,892	108.2	0	1	0
581	ERR170972	100	558	2,289,498	105.0	0	1	0
582	ERR170973	100	634	2,596,506	119.1	0	1	0
583	ERR170974	100	600	2,456,816	112.7	0	1	0
584	ERR170975	100	954	3,906,342	179.2	0	1	0
585	ERR170976	100	560	2,293,160	105.2	0	1	0
586	ERR170977	100	558	2,285,002	104.8	0	1	0
587	ERR170978	100	580	2,373,362	108.9	0	1	0
588	ERR170979	100	548	2,248,418	103.1	0	1	0
589	ERR170980	100	570	2,334,666	107.1	0	1	0
590	ERR170981	100	542	2,220,384	101.9	0	1	0
591	ERR170982	100	606	2,478,096	113.7	0	1	0
592	ERR170983	100	566	2,319,120	106.4	0	1	0
593	ERR170984	100	544	2,231,866	102.4	0	1	0
594	ERR170985	100	566	2,316,404	106.3	0	1	0
595	ERR170986	100	536	2,196,032	100.7	0	1	0
596	ERR170987	100	610	2,499,418	114.7	0	1	0
597	ERR170988	100	528	2,159,100	99.0	0	1	0
598	ERR170989	100	538	2,200,948	101.0	0	1	0
599	ERR170990	100	458	1,875,166	86.0	0	1	0
600	ERR170991	100	564	2,308,802	105.9	0	1	0
601	ERR170992	100	570	2,338,678	107.3	0	1	0
602	ERR170993	100	570	2,336,266	107.2	0	1	0
603	ERR170994	100	532	2,178,570	99.9	0	1	0
604	ERR170995	100	612	2,506,000	115.0	0	1	0
605	ERR170996	100	510	2,091,278	95.9	0	1	0
606	ERR170997	100	582	2,385,230	109.4	0	1	0
607	ERR170998	100	554	2,267,858	104.0	0	1	0
608	ERR170999	100	568	2,328,284	106.8	0	1	0
609	ERR171000	100	596	2,441,704	112.0	0	1	0
610	ERR171001	100	638	2,613,518	119.9	0	1	0
611	ERR171002	100	568	2,323,020	106.6	0	1	0
612	ERR171003	100	594	2,428,966	111.4	0	1	0
613	ERR171004	100	542	2,221,904	101.9	0	1	0
614	ERR171005	100	546	2,239,802	102.7	0	1	0
615	ERR171006	100	564	2,307,472	105.8	0	1	0
616	ERR171007	100	580	2,374,826	108.9	0	1	0
617	ERR171008	100	550	2,250,942	103.3	0	1	0
618	ERR171009	100	632	2,589,500	118.8	0	1	0
619	ERR171010	100	628	2,571,954	118.0	0	1	0
620	ERR171011	100	658	2,695,862	123.7	0	1	0
621	ERR171012	100	576	2,355,366	108.0	0	1	0
622	ERR171013	100	570	2,334,390	107.1	0	1	0
623	ERR171014	100	562	2,299,368	105.5	0	1	0
624	ERR171015	100	604	2,472,976	113.4	0	1	0
625	ERR171016	100	534	2,188,628	100.4	0	1	0
626	ERR171017	100	566	2,320,140	106.4	0	1	0
627	ERR171018	100	550	2,255,786	103.5	0	1	0
628	ERR171019	100	614	2,513,556	115.3	0	1	0
629	ERR171020	100	584	2,389,912	109.6	0	1	0
630	ERR171021	100	566	2,315,106	106.2	0	1	0
631	ERR171022	100	586	2,396,582	109.9	0	1	0
632	ERR171023	100	610	2,500,924	114.7	0	1	0
633	ERR171024	100	616	2,519,686	115.6	0	1	0
634	ERR171025	100	528	2,163,586	99.2	0	1	0
635	ERR171026	100	544	2,234,628	102.5	0	1	0
636	ERR171027	100	568	2,332,566	107.0	0	1	0
637	ERR171028	100	588	2,417,332	110.9	0	1	0
638	ERR171029	100	606	2,490,188	114.2	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
639	ERR171030	100	530	2,182,258	100.1	0	1	0
640	ERR171031	100	554	2,282,032	104.7	0	1	0
641	ERR171032	100	654	2,688,142	123.3	0	1	0
642	ERR171033	100	602	2,478,654	113.7	0	1	0
643	ERR171034	100	606	2,493,032	114.4	0	1	0
644	ERR171035	100	622	2,544,536	116.7	0	1	0
645	ERR171036	100	620	2,539,572	116.5	0	1	0
646	ERR171037	100	606	2,478,978	113.7	0	1	0
647	ERR171038	100	628	2,570,854	117.9	0	1	0
648	ERR171039	100	590	2,420,134	111.0	0	1	0
649	ERR171040	100	636	2,606,314	119.6	0	1	0
650	ERR171041	100	678	2,775,048	127.3	0	1	0
651	ERR171042	100	532	2,180,276	100.0	0	1	0
652	ERR171043	100	644	2,641,140	121.2	0	1	0
653	ERR171044	100	580	2,375,810	109.0	0	1	0
654	ERR171045	100	524	2,143,582	98.3	0	1	0
655	ERR171046	100	652	2,673,692	122.6	0	1	0
656	ERR171047	100	668	2,738,744	125.6	0	1	0
657	ERR171048	100	644	2,638,892	121.1	0	1	0
658	ERR171049	100	578	2,364,592	108.5	0	1	0
659	ERR171050	100	584	2,395,326	109.9	0	1	0
660	ERR171051	100	576	2,361,942	108.3	0	1	0
661	ERR171052	100	494	2,023,332	92.8	0	1	0
662	ERR171053	100	536	2,197,590	100.8	0	1	0
663	ERR171054	100	550	2,253,950	103.4	0	1	0
664	ERR171055	100	640	2,624,586	120.4	0	1	0
665	ERR171056	100	618	2,534,006	116.2	0	1	0
666	ERR171057	100	610	2,499,796	114.7	0	1	0
667	ERR171058	100	602	2,468,396	113.2	0	1	0
668	ERR171059	100	676	2,769,800	127.1	0	1	0
669	ERR171060	100	630	2,583,596	118.5	0	1	0
670	ERR171061	100	604	2,470,920	113.3	0	1	0
671	ERR171062	100	568	2,326,006	106.7	0	1	0
672	ERR171063	100	618	2,529,538	116.0	0	1	0
673	ERR171064	100	620	2,543,264	116.7	0	1	0
674	ERR171065	100	600	2,457,182	112.7	0	1	0
675	ERR171066	100	658	2,697,072	123.7	0	1	0
676	ERR171067	100	616	2,518,886	115.5	0	1	0
677	ERR171068	100	642	2,626,686	120.5	0	1	0
678	ERR171069	100	670	2,747,106	126.0	0	1	0
679	ERR171070	100	610	2,497,858	114.6	0	1	0
680	ERR171071	100	690	2,828,510	129.7	0	1	0
681	ERR171072	100	578	2,366,742	108.6	0	1	0
682	ERR171073	100	734	3,002,270	137.7	0	1	0
683	ERR171074	100	508	2,081,880	95.5	0	1	0
684	ERR171075	100	554	2,269,466	104.1	0	1	0
685	ERR171076	100	552	2,258,144	103.6	0	1	0
686	ERR171077	100	566	2,315,226	106.2	0	1	0
687	ERR171078	100	568	2,324,214	106.6	0	1	0
688	ERR171079	100	636	2,607,266	119.6	0	1	0
689	ERR171080	100	550	2,250,822	103.2	0	1	0
690	ERR171081	100	608	2,488,820	114.2	0	1	0
691	ERR171082	100	534	2,185,740	100.3	0	1	0
692	ERR171083	100	642	2,629,582	120.6	0	1	0
693	ERR171084	100	642	2,627,484	120.5	0	1	0
694	ERR171085	100	622	2,550,384	117.0	0	1	0
695	ERR171086	100	536	2,193,876	100.6	0	1	0
696	ERR171087	100	616	2,526,812	115.9	0	1	0
697	ERR171088	100	608	2,490,536	114.2	0	1	0
698	ERR171089	100	642	2,625,340	120.4	0	1	0
699	ERR171090	100	590	2,413,042	110.7	0	1	0
700	ERR171091	100	658	2,695,508	123.6	0	1	0
701	ERR171092	100	646	2,645,296	121.3	0	1	0
702	ERR171093	100	676	2,771,564	127.1	0	1	0
703	ERR171094	100	634	2,599,222	119.2	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
704	ERR171095	100	650	2,660,374	122.0	0	1	0
705	ERR171096	100	586	2,400,904	110.1	0	1	0
706	ERR171097	100	602	2,469,738	113.3	0	1	0
707	ERR171098	100	528	2,159,180	99.0	0	1	0
708	ERR171099	100	538	2,203,166	101.1	0	1	0
709	ERR171100	100	536	2,194,268	100.7	0	1	0
710	ERR171101	100	582	2,381,752	109.3	0	1	0
711	ERR171102	100	590	2,418,116	110.9	0	1	0
712	ERR171103	100	596	2,441,172	112.0	0	1	0
713	ERR171104	100	594	2,429,204	111.4	0	1	0
714	ERR171105	100	642	2,629,162	120.6	0	1	0
715	ERR171106	100	636	2,605,870	119.5	0	1	0
716	ERR171107	100	578	2,364,998	108.5	0	1	0
717	ERR171108	100	606	2,483,078	113.9	0	1	0
718	ERR171109	100	572	2,341,708	107.4	0	1	0
719	ERR171110	100	650	2,662,048	122.1	0	1	0
720	ERR171111	100	664	2,721,894	124.9	0	1	0
721	ERR171112	100	668	2,738,800	125.6	0	1	0
722	ERR171113	100	642	2,626,534	120.5	0	1	0
723	ERR171114	100	608	2,493,796	114.4	0	1	0
724	ERR171115	100	692	2,837,646	130.2	0	1	0
725	ERR171116	100	682	2,792,084	128.1	0	1	0
726	ERR171117	100	648	2,654,700	121.8	0	1	0
727	ERR171118	100	626	2,564,576	117.6	0	1	0
728	ERR171119	100	714	2,924,904	134.2	0	1	0
729	ERR171120	100	666	2,724,334	125.0	0	1	0
730	ERR171121	100	688	2,818,434	129.3	0	1	0
731	ERR172493	100	1,696	6,939,706	318.3	0	1	0
732	ERR172494	100	2,256	9,224,834	423.2	0	1	0
733	ERR172495	100	1,464	5,987,216	274.6	0	1	0
734	ERR172496	100	698	2,860,634	131.2	0	1	0
735	ERR172497	100	1,662	6,797,018	311.8	0	1	0
736	ERR172498	100	1,700	6,955,622	319.1	0	1	0
737	ERR172499	100	1,566	6,402,958	293.7	0	1	0
738	ERR172500	100	930	3,805,370	174.6	0	1	0
739	ERR172501	100	976	3,991,580	183.1	0	1	0
740	ERR172502	100	1,594	6,494,344	297.9	0	1	0
741	ERR172503	100	940	3,831,832	175.8	0	1	0
742	ERR172504	100	922	3,756,612	172.3	0	1	0
743	ERR172505	100	626	2,551,336	117.0	0	1	0
744	ERR172506	100	680	2,774,392	127.3	0	1	0
745	ERR172507	100	792	3,231,778	148.2	0	1	0
746	ERR172508	100	1,152	4,693,378	215.3	0	1	0
747	ERR172509	100	492	2,005,930	92.0	0	1	0
748	ERR172510	100	600	2,448,730	112.3	0	1	0
749	ERR172511	100	488	1,994,346	91.5	0	1	0
750	ERR172512	100	558	2,275,564	104.4	0	1	0
751	ERR172513	100	696	2,835,002	130.0	0	1	0
752	ERR172514	100	1,554	6,331,212	290.4	0	1	0
753	ERR172515	100	924	3,767,346	172.8	0	1	0
754	ERR172516	100	756	3,080,982	141.3	0	1	0
755	ERR172517	100	660	2,689,550	123.4	0	1	0
756	ERR172518	100	702	2,864,234	131.4	0	1	0
757	ERR172519	100	780	3,181,566	145.9	0	1	0
758	ERR172520	100	730	2,974,590	136.4	0	1	0
759	ERR172521	100	758	3,092,330	141.9	0	1	0
760	ERR172522	100	806	3,288,276	150.8	0	1	0
761	ERR172523	100	850	3,464,520	158.9	0	1	0
762	ERR172524	100	660	2,692,446	123.5	0	1	0
763	ERR172525	100	568	2,314,564	106.2	0	1	0
764	ERR172526	100	814	3,321,632	152.4	0	1	0
765	ERR172527	100	824	3,361,414	154.2	0	1	0
766	ERR172528	100	1,368	5,579,262	255.9	0	1	0
767	ERR172529	100	816	3,325,934	152.6	0	1	0
768	ERR172530	100	888	3,622,396	166.2	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
769	ERR172531	100	1,754	7,149,050	327.9	0	1	0
770	ERR172532	100	1,012	4,122,424	189.1	0	1	0
771	ERR172533	100	732	2,983,256	136.8	0	1	0
772	ERR172534	100	908	3,705,366	170.0	0	1	0
773	ERR172535	100	738	3,013,902	138.3	0	1	0
774	ERR172536	100	958	3,906,338	179.2	0	1	0
775	ERR172537	100	680	2,771,950	127.2	0	1	0
776	ERR172538	100	1,034	4,213,982	193.3	0	1	0
777	ERR172539	100	934	3,807,006	174.6	0	1	0
778	ERR172540	100	1,100	4,481,508	205.6	0	1	0
779	ERR172541	100	614	2,507,280	115.0	0	1	0
780	ERR172542	100	616	2,513,428	115.3	0	1	0
781	ERR172543	100	734	2,993,674	137.3	0	1	0
782	ERR172544	100	744	3,032,408	139.1	0	1	0
783	ERR172545	100	726	2,957,362	135.7	0	1	0
784	ERR172546	100	702	2,861,060	131.2	0	1	0
785	ERR172547	100	762	3,106,282	142.5	0	1	0
786	ERR172548	100	602	2,457,664	112.7	0	1	0
787	ERR172549	100	800	3,265,006	149.8	0	1	0
788	ERR172550	100	886	3,612,198	165.7	0	1	0
789	ERR172551	100	806	3,287,114	150.8	0	1	0
790	ERR172552	100	880	3,584,420	164.4	0	1	0
791	ERR172553	100	632	2,578,254	118.3	0	1	0
792	ERR172554	100	736	3,002,298	137.7	0	1	0
793	ERR172555	100	842	3,429,870	157.3	0	1	0
794	ERR172556	100	746	3,041,522	139.5	0	1	0
795	ERR172557	100	606	2,473,016	113.4	0	1	0
796	ERR172558	100	714	2,915,638	133.7	0	1	0
797	ERR172559	100	686	2,798,630	128.4	0	1	0
798	ERR172560	100	788	3,211,160	147.3	0	1	0
799	ERR172561	100	514	2,094,384	96.1	0	1	0
800	ERR172562	100	674	2,751,300	126.2	0	1	0
801	ERR172563	100	816	3,324,858	152.5	0	1	0
802	ERR172564	100	722	2,944,746	135.1	0	1	0
803	ERR172565	100	872	3,552,856	163.0	0	1	0
804	ERR172566	100	854	3,484,860	159.9	0	1	0
805	ERR172567	100	772	3,147,850	144.4	0	1	0
806	ERR172568	100	944	3,848,492	176.5	0	1	0
807	ERR172569	100	886	3,610,306	165.6	0	1	0
808	ERR172570	100	872	3,558,174	163.2	0	1	0
809	ERR172571	100	852	3,476,566	159.5	0	1	0
810	ERR172572	100	860	3,508,966	161.0	0	1	0
811	ERR172573	100	948	3,868,708	177.5	0	1	0
812	ERR172574	100	858	3,499,188	160.5	0	1	0
813	ERR172575	100	838	3,414,976	156.7	0	1	0
814	ERR172576	100	798	3,257,346	149.4	0	1	0
815	ERR172577	100	898	3,665,096	168.1	0	1	0
816	ERR172578	100	1,674	6,820,634	312.9	0	1	0
817	ERR172579	100	952	3,880,968	178.0	0	1	0
818	ERR172580	100	1,006	4,097,804	188.0	0	1	0
819	ERR172581	100	960	3,910,390	179.4	0	1	0
820	ERR172582	100	1,014	4,130,434	189.5	0	1	0
821	ERR172583	100	1,046	4,260,666	195.4	0	1	0
822	ERR172584	100	990	4,033,510	185.0	0	1	0
823	ERR172585	100	874	3,560,038	163.3	0	1	0
824	ERR172586	100	1,362	5,550,838	254.6	0	1	0
825	ERR172587	100	992	4,047,870	185.7	0	1	0
826	ERR172588	100	1,028	4,191,030	192.2	0	1	0
827	ERR176024	100	838	3,413,928	156.6	0	1	0
828	ERR176025	100	820	3,341,902	153.3	0	1	0
829	ERR176026	100	662	2,699,222	123.8	0	1	0
830	ERR176027	100	674	2,751,700	126.2	0	1	0
831	ERR176028	100	722	2,943,488	135.0	0	1	0
832	ERR176029	100	846	3,446,724	158.1	0	1	0
833	ERR176030	100	716	2,918,894	133.9	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
834	ERR176031	100	736	3,003,028	137.8	0	1	0
835	ERR176032	100	826	3,370,236	154.6	0	1	0
836	ERR176033	100	886	3,609,194	165.6	0	1	0
837	ERR176034	100	742	3,028,760	138.9	0	1	0
838	ERR176035	100	796	3,242,880	148.8	0	1	0
839	ERR176036	100	764	3,119,290	143.1	0	1	0
840	ERR176037	100	800	3,265,070	149.8	0	1	0
841	ERR176038	100	754	3,077,556	141.2	0	1	0
842	ERR176039	100	804	3,276,800	150.3	0	1	0
843	ERR176040	100	726	2,961,912	135.9	0	1	0
844	ERR176041	100	850	3,464,078	158.9	0	1	0
845	ERR176042	100	672	2,740,044	125.7	0	1	0
846	ERR176043	100	826	3,367,232	154.5	0	1	0
847	ERR176044	100	810	3,301,590	151.4	0	1	0
848	ERR176045	100	864	3,522,766	161.6	0	1	0
849	ERR176046	100	776	3,166,632	145.3	0	1	0
850	ERR176047	100	742	3,026,746	138.8	0	1	0
851	ERR176048	100	818	3,336,074	153.0	0	1	0
852	ERR176049	100	862	3,511,698	161.1	0	1	0
853	ERR1994579	150	734	872,765	57.6	1	0	1
854	ERR1994623	150	860	1,022,969	67.5	0	0	1
855	ERR1994645	150	639	760,066	50.2	0	0	1
856	ERR1994653	150	473	562,839	37.2	1	0	1
857	ERR2113642	150	1,620	1,921,990	126.9	0	0	1
858	ERR2113664	150	1,329	1,577,688	104.1	1	0	1
859	ERR2258952	150	1,146	1,592,666	105.1	1	0	1
860	ERR2258988	150	1,151	1,599,964	105.6	0	0	1
861	ERR2259004	150	1,083	1,505,845	99.4	1	0	1
862	ERR2259007	150	1,008	1,402,883	92.6	0	0	1
863	ERR2259019	150	1,031	1,433,936	94.7	1	0	1
864	ERR2259054	150	967	1,345,974	88.8	1	0	1
865	ERR237214	100	544	2,461,506	112.9	0	1	0
866	ERR237215	100	908	4,097,532	188.0	0	1	0
867	ERR237216	100	670	3,031,790	139.1	0	1	0
868	ERR237217	100	556	2,519,026	115.6	0	1	0
869	ERR237218	100	766	3,461,946	158.8	0	1	0
870	ERR237219	100	400	1,809,498	83.0	0	1	0
871	ERR237222	100	682	3,086,722	141.6	0	1	0
872	ERR237223	100	1,290	5,819,498	266.9	0	1	0
873	ERR237224	100	1,724	7,777,370	356.8	0	1	0
874	ERR237225	100	904	4,084,086	187.3	0	1	0
875	ERR237226	100	984	4,439,102	203.6	0	1	0
876	ERR237230	100	808	3,295,284	151.2	0	1	0
877	ERR237231	100	560	2,531,196	116.1	0	1	0
878	ERR237232	100	390	1,767,032	81.1	0	1	0
879	ERR237233	100	968	4,373,626	200.6	0	1	0
880	ERR237234	100	428	1,940,654	89.0	0	1	0
881	ERR237238	100	468	2,121,336	97.3	0	1	0
882	ERR237239	100	2,212	9,972,286	457.4	0	1	0
883	ERR237240	100	638	2,889,020	132.5	0	1	0
884	ERR237241	100	704	3,183,660	146.0	0	1	0
885	ERR237242	100	1,286	5,800,322	266.1	0	1	0
886	ERR237246	100	724	3,275,454	150.3	0	1	0
887	ERR237247	100	454	2,058,454	94.4	0	1	0
888	ERR237248	100	930	4,203,864	192.8	0	1	0
889	ERR237249	100	734	3,317,132	152.2	0	1	0
890	ERR237250	100	928	4,192,460	192.3	0	1	0
891	ERR237254	100	836	3,778,018	173.3	0	1	0
892	ERR237255	100	2,108	9,502,944	435.9	0	1	0
893	ERR237256	100	464	2,100,702	96.4	0	1	0
894	ERR237257	100	1,358	6,126,572	281.0	0	1	0
895	ERR237258	100	1,374	6,200,608	284.4	0	1	0
896	ERR237288	100	506	2,291,782	105.1	0	1	0
897	ERR237289	100	1,496	6,743,482	309.3	0	1	0
898	ERR237290	100	1,228	5,544,384	254.3	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
899	ERR237291	100	906	4,091,674	187.7	0	1	0
900	ERR237292	100	910	4,114,250	188.7	0	1	0
901	ERR237293	100	1,140	5,144,614	236.0	0	1	0
902	ERR237294	100	1,022	4,609,954	211.5	0	1	0
903	ERR237295	100	312	1,416,140	65.0	0	1	0
904	ERR237296	100	460	2,085,178	95.7	0	1	0
905	ERR237297	100	764	3,451,028	158.3	0	1	0
906	ERR237298	100	418	1,893,054	86.8	0	1	0
907	ERR237299	100	462	2,092,932	96.0	0	1	0
908	ERR237300	100	440	1,998,156	91.7	0	1	0
909	ERR237303	100	1,102	4,973,328	228.1	0	1	0
910	ERR237304	100	646	2,920,500	134.0	0	1	0
911	ERR237305	100	504	2,279,886	104.6	0	1	0
912	ERR237306	100	466	2,107,742	96.7	0	1	0
913	ERR237307	100	656	2,962,454	135.9	0	1	0
914	ERR237311	100	516	2,332,634	107.0	0	1	0
915	ERR237312	100	398	1,802,530	82.7	0	1	0
916	ERR237313	100	438	1,987,370	91.2	0	1	0
917	ERR237314	100	368	1,664,470	76.4	0	1	0
918	ERR237315	100	768	3,470,112	159.2	0	1	0
919	ERR237316	100	1,210	5,462,054	250.6	0	1	0
920	ERR237319	100	766	3,462,514	158.8	0	1	0
921	ERR237320	100	448	2,034,182	93.3	0	1	0
922	ERR237321	100	434	1,964,690	90.1	0	1	0
923	ERR237322	100	540	2,444,000	112.1	0	1	0
924	ERR237323	100	1,514	6,830,626	313.3	0	1	0
925	ERR237324	100	486	2,205,088	101.2	0	1	0
926	ERR237327	100	506	2,287,506	104.9	0	1	0
927	ERR237328	100	330	1,499,072	68.8	0	1	0
928	ERR237329	100	432	1,956,256	89.7	0	1	0
929	ERR237330	100	382	1,735,346	79.6	0	1	0
930	ERR237331	100	430	1,947,862	89.4	0	1	0
931	ERR237332	100	668	3,018,232	138.5	0	1	0
932	ERR237335	100	438	1,981,964	90.9	0	1	0
933	ERR237336	100	374	1,692,930	77.7	0	1	0
934	ERR237337	100	508	2,297,808	105.4	0	1	0
935	ERR237338	100	362	1,637,158	75.1	0	1	0
936	ERR237339	100	736	3,323,262	152.4	0	1	0
937	ERR237340	100	564	2,557,086	117.3	0	1	0
938	ERR237342	100	584	2,641,862	121.2	0	1	0
939	ERR2619334	150	705	838,423	55.3	0	0	1
940	ERR2619490	150	578	687,798	45.4	1	0	1
941	ERR2619502	150	597	709,730	46.8	0	0	1
942	ERR2619551	150	670	796,858	52.6	0	0	1
943	ERR2619556	150	592	703,655	46.4	1	0	1
944	ERR2619561	150	535	636,746	42.0	0	0	1
945	ERR310519	100	1,014	4,574,974	209.9	0	1	0
946	ERR310523	100	932	4,207,466	193.0	0	1	0
947	ERR310524	100	922	4,167,272	191.2	0	1	0
948	ERR310525	100	964	4,356,886	199.9	0	1	0
949	ERR310526	100	966	4,359,786	200.0	0	1	0
950	ERR310530	100	988	4,459,498	204.6	0	1	0
951	ERR310532	100	1,034	4,668,976	214.2	0	1	0
952	ERR310534	100	1,012	4,569,384	209.6	0	1	0
953	ERR310537	100	1,020	4,600,994	211.1	0	1	0
954	ERR310538	100	968	4,368,556	200.4	0	1	0
955	ERR310539	100	960	4,333,914	198.8	0	1	0
956	ERR310540	100	1,040	4,694,900	215.4	0	1	0
957	ERR310541	100	1,006	4,540,732	208.3	0	1	0
958	ERR310542	100	1,036	4,673,102	214.4	0	1	0
959	ERR310543	100	1,054	4,758,224	218.3	0	1	0
960	ERR314084	100	1,194	5,384,556	247.0	0	1	0
961	ERR314085	100	598	2,706,672	124.2	0	1	0
962	ERR314086	100	1,276	5,758,990	264.2	0	1	0
963	ERR314087	100	1,050	4,740,446	217.5	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
964	ERR314088	100	918	4,149,906	190.4	0	1	0
965	ERR314089	100	976	4,406,718	202.1	0	1	0
966	ERR314090	100	988	4,462,356	204.7	0	1	0
967	ERR314091	100	990	4,473,102	205.2	0	1	0
968	ERR314092	100	1,318	5,944,966	272.7	0	1	0
969	ERR314093	100	818	3,698,396	169.7	0	1	0
970	ERR314094	100	1,132	5,104,460	234.1	0	1	0
971	ERR314095	100	1,070	4,830,402	221.6	0	1	0
972	ERR314096	100	1,010	4,558,614	209.1	0	1	0
973	ERR314097	100	902	4,076,162	187.0	0	1	0
974	ERR314098	100	1,108	5,000,488	229.4	0	1	0
975	ERR314099	100	1,096	4,944,134	226.8	0	1	0
976	ERR314100	100	1,240	5,592,354	256.5	0	1	0
977	ERR314101	100	852	3,846,174	176.4	0	1	0
978	ERR314102	100	1,156	5,214,812	239.2	0	1	0
979	ERR314103	100	1,116	5,032,748	230.9	0	1	0
980	ERR314104	100	776	3,509,804	161.0	0	1	0
981	ERR314105	100	904	4,086,296	187.4	0	1	0
982	ERR314106	100	1,014	4,574,586	209.8	0	1	0
983	ERR314107	100	1,074	4,849,312	222.4	0	1	0
984	ERR314108	100	1,206	5,442,768	249.7	0	1	0
985	ERR314109	100	872	3,943,220	180.9	0	1	0
986	ERR314110	100	1,026	4,633,126	212.5	0	1	0
987	ERR314111	100	1,110	5,009,758	229.8	0	1	0
988	ERR314112	100	974	4,399,038	201.8	0	1	0
989	ERR314113	100	896	4,045,172	185.6	0	1	0
990	ERR314114	100	976	4,405,702	202.1	0	1	0
991	ERR314115	100	968	4,372,386	200.6	0	1	0
992	ERR314116	100	1,334	6,021,350	276.2	0	1	0
993	ERR314117	100	890	4,020,824	184.4	0	1	0
994	ERR314118	100	1,110	5,006,616	229.7	0	1	0
995	ERR314119	100	1,080	4,870,948	223.4	0	1	0
996	ERR314120	100	754	3,408,816	156.4	0	1	0
997	ERR314121	100	772	3,487,142	160.0	0	1	0
998	ERR314122	100	944	4,265,516	195.7	0	1	0
999	ERR314123	100	1,074	4,851,450	222.5	0	1	0
1000	ERR314124	100	1,144	5,162,582	236.8	0	1	0
1001	ERR314125	100	910	4,109,064	188.5	0	1	0
1002	ERR314126	100	962	4,345,192	199.3	0	1	0
1003	ERR314127	100	958	4,328,592	198.6	0	1	0
1004	ERR314128	100	876	3,954,560	181.4	0	1	0
1005	ERR314129	100	1,084	4,888,994	224.3	0	1	0
1006	ERR314130	100	976	4,403,642	202.0	0	1	0
1007	ERR314131	100	1,026	4,629,822	212.4	0	1	0
1008	ERR314132	100	1,290	5,824,104	267.2	0	1	0
1009	ERR314133	100	916	4,139,420	189.9	0	1	0
1010	ERR314134	100	922	4,167,050	191.1	0	1	0
1011	ERR314135	100	1,056	4,766,052	218.6	0	1	0
1012	ERR314136	100	836	3,779,716	173.4	0	1	0
1013	ERR314137	100	1,060	4,784,458	219.5	0	1	0
1014	ERR314138	100	988	4,459,410	204.6	0	1	0
1015	ERR314139	100	1,008	4,554,132	208.9	0	1	0
1016	ERR314140	100	1,298	5,859,634	268.8	0	1	0
1017	ERR314141	100	984	4,446,430	204.0	0	1	0
1018	ERR314142	100	822	3,712,636	170.3	0	1	0
1019	ERR314143	100	1,116	5,034,252	230.9	0	1	0
1020	ERR314144	100	1,052	4,746,858	217.7	0	1	0
1021	ERR314145	100	1,192	5,378,030	246.7	0	1	0
1022	ERR314146	100	1,088	4,910,616	225.3	0	1	0
1023	ERR314147	100	1,050	4,741,710	217.5	0	1	0
1024	ERR314148	100	1,132	5,109,708	234.4	0	1	0
1025	ERR314149	100	1,024	4,620,628	212.0	0	1	0
1026	ERR314150	100	996	4,497,982	206.3	0	1	0
1027	ERR314151	100	1,034	4,667,654	214.1	0	1	0
1028	ERR314152	100	794	3,584,314	164.4	0	1	0

Table 14. Continued.

#	ENA/SRA Accn. Number	Avg. Read Length (bp)	Sample Size		Sequence Depth (x)	Included in Test (1=included, 0=not included)		
			MB	Reads		MLST Comparison	Large-scale MLST Accuracy	Larger Schemes
1029	ERR314153	100	1,100	4,967,082	227.8	0	1	0
1030	ERR314154	100	936	4,225,154	193.8	0	1	0
1031	ERR314155	100	1,168	5,274,240	241.9	0	1	0
1032	ERR314156	100	1,418	6,396,328	293.4	0	1	0
1033	ERR314157	100	1,078	4,869,432	223.4	0	1	0
1034	ERR314158	100	1,098	4,954,166	227.3	0	1	0
1035	ERR314159	100	1,036	4,681,406	214.7	0	1	0
1036	ERR314160	100	770	3,479,360	159.6	0	1	0
1037	ERR557644	151	1,988	5,600,448	387.9	0	1	0
1038	ERR957622	151	1,460	4,112,856	284.9	0	1	0
1039	ERR977559	151	1,443	1,695,282	112.7	0	0	1
1040	SRR7352647	290	1,138	909,257	116.0	1	0	1
S. pneumoniae – Genome size: **2,038,615 bp; Sequencing platform: Illumina**								
1041	ERR226388	100	868	3,304,660	162.1	1	0	0
1042	ERR226389	100	699	2,667,968	130.9	1	0	0
1043	ERR388926	100	920	3,501,988	171.8	1	0	0
1044	ERR470408	100	1,210	4,594,698	225.4	1	0	0
1045	ERR697245	100	1,232	4,679,456	229.5	1	0	0
1046	ERR755310	100	1,375	5,217,458	255.9	1	0	0
1047	ERR755504	100	1,597	6,057,210	297.1	1	0	0
1048	ERR755581	100	1,521	5,769,448	283.0	1	0	0
1049	ERR755613	100	1,486	5,636,538	276.5	1	0	0
1050	ERR755622	100	1,244	4,723,332	231.7	1	0	0

Table 15. Assemblies used for the limit of detection, and single and multithread performance tests.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1	GCA_001224405	Campylobacter jejuni	OXC6519	19270	GCA_001224405.1	1.70287	Scaffold	21	2	1	1	3	2	1	5
2	GCA_001226425	Campylobacter jejuni	OXC6417	19270	GCA_001226425.1	1.66357	Contig	42	1	2	3	4	5	9	3
3	GCA_001232325	Campylobacter jejuni	OXC6430	19270	GCA_001232325.1	1.70481	Scaffold	572	62	4	5	2	2	1	5
4	GCA_001232565	Campylobacter jejuni	OXC6552	19270	GCA_001232565.1	1.69542	Scaffold	3769	2	1	12	3	2	1	12
5	GCA_001238785	Campylobacter jejuni	OXC6542	19270	GCA_001238785.1	1.83943	Scaffold	573	7	28	4	28	17	34	12
6	GCA_001406915	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001406915.1	1.60671	Scaffold	572	62	4	5	2	2	1	5
7	GCA_001407155	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407155.1 (contaminated)	4.30156	Scaffold	464	24	2	2	2	10	3	1
8	GCA_001407175	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407175.1 (contaminated)	4.13204	Scaffold	464	24	2	2	2	10	3	1
9	GCA_001407215	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407215.1 (contaminated)	9.12363	Scaffold	42	1	2	3	4	5	9	3
10	GCA_001407395	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407395.1 (contaminated)	4.35099	Scaffold	607	8	2	5	53	11	3	1
11	GCA_001407875	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407875.1 (contaminated)	4.38242	Scaffold	2364	14	249	5	2	11	3	6
12	GCA_001488115	Campylobacter jejuni	n/a	19270	GCA_001488115.1	1.64043	Contig	48	2	4	1	2	7	1	5
13	GCA_001489095	Campylobacter jejuni	n/a	19270	GCA_001489095.1	1.73999	Contig	262	2	1	1	3	2	1	3
14	GCA_001490635	Campylobacter jejuni	n/a	19270	GCA_001490635.1	1.63286	Contig	677	10	81	50	99	120	76	52
15	GCA_001491455	Campylobacter jejuni	n/a	19270	GCA_001491455.1	1.68551	Contig	574	7	53	2	10	11	3	3
16	GCA_001494015	Campylobacter jejuni	n/a	19270	GCA_001494015.1	1.73246	Contig	262	2	1	1	3	2	1	3
17	GCA_001494515	Campylobacter jejuni	n/a	19270	GCA_001494515.1	1.64326	Contig	257	9	2	4	62	4	5	6
18	GCA_001496075	Campylobacter jejuni	n/a	19270	GCA_001496075.1	1.64829	Contig	257	9	2	4	62	4	5	6
19	GCA_001498975	Campylobacter jejuni	n/a	19270	GCA_001498975.1	1.70409	Contig	45	4	7	10	4	1	7	1
20	GCA_001506325	Campylobacter jejuni	CJ677CC010	19270	GCA_001506325.1	1.65386	Complete Genome	677	10	81	50	99	120	76	52
21	GCF_000184805	Campylobacter jejuni subsp. jejuni DFVF1099	DFVF1099	19270	GCA_000184805.2	1.73386	Contig	21	2	1	1	3	2	1	5
22	GCF_000184845	Campylobacter jejuni subsp. jejuni 327	327	19270	GCA_000184845.2	1.61861	Contig	230	4	7	41	4	42	7	1
23	GCF_000234525	Campylobacter jejuni subsp. jejuni NW	NW	19270	GCA_000234525.1	1.6527	Scaffold	354	8	10	2	2	11	12	6
24	GCF_000234545	Campylobacter jejuni subsp. jejuni D2600	D2600	19270	GCA_000234545.1	1.62241	Scaffold	429	7	4	5	2	11	1	5
25	GCF_000242375	Campylobacter jejuni subsp. jejuni P110B	P110B	19270	GCA_000242375.2	1.65634	Contig	474	2	4	1	2	2	1	5
26	GCF_000242395	Campylobacter jejuni subsp. jejuni H22082	H22082	19270	GCA_000242395.2	1.65912	Contig	474	2	4	1	2	2	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
27	GCF_000251165	Campylobacter jejuni subsp. jejuni ATCC 33560	ATCC 33560	19270	GCA_000251165.2	1.73219	Contig	403	10	27	16	19	10	5	7
28	GCF_000254255	Campylobacter jejuni subsp. jejuni 129-258	129-258	19270	GCA_000254255.2	1.60816	Contig	459	1	2	3	3	5	9	3
29	GCF_000254275	Campylobacter jejuni subsp. jejuni 51494	51494	19270	GCA_000254275.2	1.80364	Contig	4834	103	2	5	2	156	3	6
30	GCF_000254295	Campylobacter jejuni subsp. jejuni LMG 23216	LMG 23216	19270	GCA_000254295.2	1.47429	Contig	4835	64	89	319	100	94	103	16
31	GCF_000254315	Campylobacter jejuni subsp. jejuni LMG 23218	LMG 23218	19270	GCA_000254315.2	1.67893	Contig	48	2	4	1	2	7	1	5
32	GCF_000254335	Campylobacter jejuni subsp. jejuni LMG 23223	LMG 23223	19270	GCA_000254335.2	1.6325	Contig	791	7	97	5	2	135	68	26
33	GCF_000254355	Campylobacter jejuni subsp. jejuni LMG 23263	LMG 23263	19270	GCA_000254355.2	1.74471	Contig	3504	7	55	5	10	11	68	6
34	GCF_000254375	Campylobacter jejuni subsp. jejuni 60004	60004	19270	GCA_000254375.2	1.67541	Contig	4836	2	378	27	2	11	3	5
35	GCF_000254395	Campylobacter jejuni subsp. jejuni LMG 23264	LMG 23264	19270	GCA_000254395.2	1.72085	Contig	46	2	21	5	3	2	1	5
36	GCF_000254415	Campylobacter jejuni subsp. jejuni LMG 23269	LMG 23269	19270	GCA_000254415.2	1.6284	Contig	4837	8	17	5	3	10	59	6
37	GCF_000254435	Campylobacter jejuni subsp. jejuni 55037	55037	19270	GCA_000254435.2	1.58972	Contig	45	4	7	10	4	1	7	1
38	GCF_000254455	Campylobacter jejuni subsp. jejuni LMG 9879	LMG 9879	19270	GCA_000254455.2	1.65081	Contig	47	2	1	1	5	2	1	5
39	GCF_000254475	Campylobacter jejuni subsp. jejuni 86605	86605	19270	GCA_000254475.2	1.6384	Contig	4840	2	4	27	122	11	1	5
40	GCF_000254495	Campylobacter jejuni subsp. jejuni LMG 23357	LMG 23357	19270	GCA_000254495.2	1.67286	Contig	4883	27	33	22	49	101	9	31
41	GCF_000254515	Campylobacter jejuni subsp. jejuni ATCC 33560	ATCC 33560	19270	GCA_000254515.2	1.71246	Contig	403	10	27	16	19	10	5	7
42	GCF_000254535	Campylobacter jejuni subsp. jejuni LMG 9081	LMG 9081	19270	GCA_000254535.2	1.59384	Contig	52	9	25	2	10	22	3	6
43	GCF_000254555	Campylobacter jejuni subsp. jejuni 53161	53161	19270	GCA_000254555.2	1.56274	Contig	4838	7	17	5	68	11	3	6
44	GCF_000254575	Campylobacter jejuni subsp. jejuni LMG 9217	LMG 9217	19270	GCA_000254575.2	1.65345	Contig	443	24	17	2	15	23	3	12
45	GCF_000254595	Campylobacter jejuni subsp. jejuni 2008-1025	2008-1025	19270	GCA_000254595.2	1.66104	Contig	50	2	1	12	3	2	1	5
46	GCF_000254615	Campylobacter jejuni subsp. jejuni 2008-894	2008-894	19270	GCA_000254615.2	1.62793	Contig	1962	55	172	21	49	125	83	51
47	GCF_000254635	Campylobacter jejuni subsp. jejuni 2008-872	2008-872	19270	GCA_000254635.2	1.60231	Contig	61	1	4	2	2	6	3	17
48	GCF_000254655	Campylobacter jejuni subsp. jejuni 2008-988	2008-988	19270	GCA_000254655.2	1.82078	Contig	572	62	4	5	2	2	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
49	GCF_000254675	Campylobacter jejuni subsp. jejuni 1997-1	1997-1	19270	GCA_000254675.2	1.6043	Contig	658	2	4	2	4	19	3	6
50	GCF_000254695	Campylobacter jejuni subsp. jejuni 2008-979	2008-979	19270	GCA_000254695.2	1.79843	Contig	2274	9	17	5	10	350	3	3
51	GCF_000254715	Campylobacter jejuni subsp. jejuni 2008-831	2008-831	19270	GCA_000254715.2	1.60941	Contig	50	2	1	12	3	2	1	5
52	GCF_000254735	Campylobacter jejuni subsp. jejuni 1997-4	1997-4	19270	GCA_000254735.2	1.67022	Contig	475	2	4	1	4	19	62	5
53	GCF_000254755	Campylobacter jejuni subsp. jejuni 1997-7	1997-7	19270	GCA_000254755.2	1.58976	Contig	61	1	4	2	2	6	3	17
54	GCF_000254775	Campylobacter jejuni subsp. jejuni 1997-10	1997-10	19270	GCA_000254775.2	1.78075	Contig	4839	9	17	2	2	86	3	309
55	GCF_000254795	Campylobacter jejuni subsp. jejuni 1997-11	1997-11	19270	GCA_000254795.2	1.60176	Contig	22	1	3	6	4	3	3	3
56	GCF_000254815	Campylobacter jejuni subsp. jejuni 1997-14	1997-14	19270	GCA_000254815.2	1.76694	Contig	5159	7	17	5	2	167	457	6
57	GCF_000254835	Campylobacter jejuni subsp. jejuni 51037	51037	19270	GCA_000254835.2	1.74736	Contig	939	7	2	5	2	156	3	6
58	GCF_000254855	Campylobacter jejuni subsp. jejuni 110-21	110-21	19270	GCA_000254855.2	1.61923	Contig	982	2	1	2	3	2	1	5
59	GCF_000254875	Campylobacter jejuni subsp. jejuni 87330	87330	19270	GCA_000254875.2	1.6104	Contig	50	2	1	12	3	2	1	5
60	GCF_000254895	Campylobacter jejuni subsp. jejuni 87459	87459	19270	GCA_000254895.2	1.728	Contig	452	7	17	12	2	10	3	6
61	GCF_000254915	Campylobacter jejuni subsp. jejuni 140-16	140-16	19270	GCA_000254915.2	1.6782	Contig	5161	1	4	2	2	225	1	17
62	GCF_000254935	Campylobacter jejuni subsp. jejuni 1213	1213	19270	GCA_000254935.2	1.64806	Contig	132	1	6	22	24	12	28	1
63	GCF_000254955	Campylobacter jejuni subsp. jejuni 1577	1577	19270	GCA_000254955.2	1.70366	Contig	122	6	4	5	2	2	1	5
64	GCF_000254975	Campylobacter jejuni subsp. jejuni 1798	1798	19270	GCA_000254975.2	1.60281	Contig	61	1	4	2	2	6	3	17
65	GCF_000254995	Campylobacter jejuni subsp. jejuni 1854	1854	19270	GCA_000254995.2	1.62251	Contig	922	1	1	2	83	2	3	6
66	GCF_000255015	Campylobacter jejuni subsp. jejuni 1893	1893	19270	GCA_000255015.2	1.62643	Contig	38	2	4	2	2	6	1	5
67	GCF_000255035	Campylobacter jejuni subsp. jejuni 1928	1928	19270	GCA_000255035.2	1.6374	Contig	806	2	1	1	3	140	3	5
68	GCF_000255055	Campylobacter jejuni subsp. jejuni LMG 9872	LMG 9872	19270	GCA_000255055.2	1.6204	Contig	677	10	81	50	99	120	76	52
69	GCF_000255075	Campylobacter jejuni subsp. jejuni LMG 23210	LMG 23210	19270	GCA_000255075.2	1.7641	Contig	380	7	2	6	10	78	37	1
70	GCF_000255095	Campylobacter jejuni subsp. jejuni LMG 23211	LMG 23211	19270	GCA_000255095.2	1.67236	Contig	220	1	6	29	2	40	32	3

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
71	GCF_000283215	Campylobacter jejuni subsp. jejuni P854	P854	19270	GCA_000283215.1	1.74727	Scaffold	573	7	28	4	28	17	34	12
72	GCF_000285675	Campylobacter jejuni subsp. jejuni xy259	xy259	19270	GCA_000285675.1	1.70978	Contig	21	2	1	1	3	2	1	5
73	GCF_000285695	Campylobacter jejuni subsp. jejuni RB922	RB922	19270	GCA_000285695.1	1.71532	Contig	21	2	1	1	3	2	1	5
74	GCF_000285715	Campylobacter jejuni subsp. jejuni 6399	6399	19270	GCA_000285715.1	1.62703	Contig	21	2	1	1	3	2	1	5
75	GCF_000285735	Campylobacter jejuni subsp. jejuni 04197	4197	19270	GCA_000285735.1	1.70134	Contig	21	2	1	1	3	2	1	5
76	GCF_000285755	Campylobacter jejuni subsp. jejuni 04199	4199	19270	GCA_000285755.1	1.73222	Contig	21	2	1	1	3	2	1	5
77	GCF_000302555	Campylobacter jejuni subsp. jejuni PT14	PT14	19270	GCA_000302555.4	1.6353	Complete Genome	50	2	1	12	3	2	1	5
78	GCF_000314085	Campylobacter jejuni subsp. jejuni BIGS0004	BIGS0004	19270	GCA_000314085.1	1.59697	Contig	45	4	7	10	4	1	7	1
79	GCF_000314265	Campylobacter jejuni subsp. jejuni BIGS0013	BIGS0013	19270	GCA_000314265.1	1.55135	Contig	61	1	4	2	2	6	3	17
80	GCF_000314285	Campylobacter jejuni subsp. jejuni BIGS0014	BIGS0014	19270	GCA_000314285.1	1.54388	Contig	2381	175	251	216	282	359	293	102
81	GCF_000314445	Campylobacter jejuni subsp. jejuni BIGS0022	BIGS0022	19270	GCA_000314445.1	1.57374	Contig	257	9	2	4	62	4	5	6
82	GCF_000355825	Campylobacter jejuni subsp. jejuni ICDCCJ07002	ICDCCJ07002	19270	GCA_000355825.1	1.69841	Scaffold	2993	1	2	42	4	11	9	8
83	GCF_000355845	Campylobacter jejuni subsp. jejuni ICDCCJ07004	ICDCCJ07004	19270	GCA_000355845.1	1.69708	Scaffold	2993	1	2	42	4	11	9	8
84	GCF_001224205	Campylobacter jejuni	OXC6262	19270	GCA_001224205.1	1.78534	Contig	904	24	2	5	53	23	3	1
85	GCF_001224265	Campylobacter jejuni	OXC6266	19270	GCA_001224265.1	1.67596	Contig	50	2	1	12	3	2	1	5
86	GCF_001224325	Campylobacter jejuni	OXC6596	19270	GCA_001224325.1	1.70537	Contig	21	2	1	1	3	2	1	5
87	GCF_001224345	Campylobacter jejuni	OXC6252	19270	GCA_001224345.1	1.678	Contig	572	62	4	5	2	2	1	5
88	GCF_001224385	Campylobacter jejuni	OXC6609	19270	GCA_001224385.1	1.63923	Contig	51	7	17	2	15	23	3	12
89	GCF_001224465	Campylobacter jejuni	OXC6300	19270	GCA_001224465.1	1.72066	Contig	2135	8	1	6	3	2	1	12
90	GCF_001224485	Campylobacter jejuni	OXC6622	19270	GCA_001224485.1	1.74184	Contig	595	7	2	1	2	10	3	6
91	GCF_001224585	Campylobacter jejuni	OXC6328	19270	GCA_001224585.1	1.69361	Contig	48	2	4	1	2	7	1	5
92	GCF_001224625	Campylobacter jejuni	OXC6640	19270	GCA_001224625.1	1.83268	Contig	464	24	2	2	2	10	3	1
93	GCF_001224665	Campylobacter jejuni	OXC6278	19270	GCA_001224665.1	1.60917	Contig	45	4	7	10	4	1	7	1
94	GCF_001224805	Campylobacter jejuni	OXC6554	19270	GCA_001224805.1	1.83827	Contig	464	24	2	2	2	10	3	1
95	GCF_001224845	Campylobacter jejuni	OXC6287	19270	GCA_001224845.1	1.70786	Contig	5718	8	10	2	2	11	487	6
96	GCF_001224865	Campylobacter jejuni	OXC6334	19270	GCA_001224865.1	1.74441	Contig	21	2	1	1	3	2	1	5
97	GCF_001224905	Campylobacter jejuni	OXC6356	19270	GCA_001224905.1	1.73754	Contig	464	24	2	2	2	10	3	1

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
98	GCF_001224965	Campylobacter jejuni	OXC6598	19270	GCA_001224965.1	1.663	Contig	50	2	1	12	3	2	1	5
99	GCF_001224985	Campylobacter jejuni	OXC6583	19270	GCA_001224985.1	1.72968	Contig	257	9	2	4	62	4	5	6
100	GCF_001225005	Campylobacter jejuni	OXC6497	19270	GCA_001225005.1	1.72379	Contig	5	7	2	5	2	10	3	6
101	GCF_001225025	Campylobacter jejuni	OXC6628	19270	GCA_001225025.1	1.63597	Contig	441	7	1	2	83	2	3	6
102	GCF_001225045	Campylobacter jejuni	OXC6345	19270	GCA_001225045.1	1.69509	Contig	48	2	4	1	2	7	1	5
103	GCF_001225065	Campylobacter jejuni	OXC6358	19270	GCA_001225065.1	1.63119	Contig	233	2	7	10	4	1	7	1
104	GCF_001225105	Campylobacter jejuni	OXC6532	19270	GCA_001225105.1	1.66954	Contig	273	2	21	5	37	60	1	5
105	GCF_001225165	Campylobacter jejuni	OXC6496	19270	GCA_001225165.1	1.70141	Contig	21	2	1	1	3	2	1	5
106	GCF_001225205	Campylobacter jejuni	OXC6436	19270	GCA_001225205.1	1.65046	Contig	1044	2	10	2	4	19	3	6
107	GCF_001225265	Campylobacter jejuni	OXC6498	19270	GCA_001225265.1	1.72431	Contig	1709	22	15	4	64	74	25	23
108	GCF_001225325	Campylobacter jejuni	OXC6265	19270	GCA_001225325.1	1.68648	Contig	48	2	4	1	2	7	1	5
109	GCF_001225345	Campylobacter jejuni	OXC6626	19270	GCA_001225345.1	1.61647	Contig	22	1	3	6	4	3	3	3
110	GCF_001225385	Campylobacter jejuni	OXC6260	19270	GCA_001225385.1	1.72761	Contig	5	7	2	5	2	10	3	6
111	GCF_001225405	Campylobacter jejuni	OXC6633	19270	GCA_001225405.1	1.69379	Contig	53	2	1	21	3	2	1	5
112	GCF_001225465	Campylobacter jejuni	OXC6319	19270	GCA_001225465.1	1.71076	Contig	356	14	17	5	2	11	3	6
113	GCF_001225565	Campylobacter jejuni	OXC6377	19270	GCA_001225565.1	1.69462	Contig	48	2	4	1	2	7	1	5
114	GCF_001225645	Campylobacter jejuni	OXC6566	19270	GCA_001225645.1	1.7306	Contig	5	7	2	5	2	10	3	6
115	GCF_001225685	Campylobacter jejuni	OXC6453	19270	GCA_001225685.1	1.6948	Contig	61	1	4	2	2	6	3	17
116	GCF_001225745	Campylobacter jejuni	OXC6509	19270	GCA_001225745.1	1.70462	Contig	1040	7	84	1	10	11	3	6
117	GCF_001225785	Campylobacter jejuni	OXC6421	19270	GCA_001225785.1	1.72333	Contig	48	2	4	1	2	7	1	5
118	GCF_001225805	Campylobacter jejuni	OXC6254	19270	GCA_001225805.1	1.71124	Contig	257	9	2	4	62	4	5	6
119	GCF_001225845	Campylobacter jejuni	OXC6306	19270	GCA_001225845.1	1.61254	Contig	61	1	4	2	2	6	3	17
120	GCF_001225885	Campylobacter jejuni	OXC6508	19270	GCA_001225885.1	1.66565	Contig	21	2	1	1	3	2	1	5
121	GCF_001225925	Campylobacter jejuni	OXC6520	19270	GCA_001225925.1	1.68714	Contig	5	7	2	5	2	10	3	6
122	GCF_001226005	Campylobacter jejuni	OXC6302	19270	GCA_001226005.1	1.63808	Contig	51	7	17	2	15	23	3	12
123	GCF_001226025	Campylobacter jejuni	OXC6320	19270	GCA_001226025.1	1.72695	Contig	986	91	2	42	4	169	9	8
124	GCF_001226045	Campylobacter jejuni	OXC6333	19270	GCA_001226045.1	1.62682	Contig	42	1	2	3	4	5	9	3
125	GCF_001226065	Campylobacter jejuni	OXC6366	19270	GCA_001226065.1	1.71015	Contig	51	7	17	2	15	23	3	12
126	GCF_001226085	Campylobacter jejuni	OXC6529	19270	GCA_001226085.1	1.70003	Scaffold	354	8	10	2	2	11	12	6
127	GCF_001226145	Campylobacter jejuni	OXC6414	19270	GCA_001226145.1	1.64654	Contig	22	1	3	6	4	3	3	3
128	GCF_001226225	Campylobacter jejuni	OXC6620	19270	GCA_001226225.1	1.67604	Contig	51	7	17	2	15	23	3	12
129	GCF_001226245	Campylobacter jejuni	OXC6289	19270	GCA_001226245.1	1.70094	Contig	50	2	1	12	3	2	1	5
130	GCF_001226345	Campylobacter jejuni	OXC6614	19270	GCA_001226345.1	1.64026	Contig	233	2	7	10	4	1	7	1
131	GCF_001226385	Campylobacter jejuni	OXC6307	19270	GCA_001226385.1	1.70481	Contig	273	2	21	5	37	60	1	5
132	GCF_001226485	Campylobacter jejuni	OXC6538	19270	GCA_001226485.1	1.64892	Contig	19	2	1	5	3	2	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
133	GCF_001226545	Campylobacter jejuni	OXC6536	19270	GCA_001226545.1	1.65566	Contig	45	4	7	10	4	1	7	1
134	GCF_001226665	Campylobacter jejuni	OXC6355	19270	GCA_001226665.1	1.66391	Contig	19	2	1	5	3	2	1	5
135	GCF_001226745	Campylobacter jejuni	OXC6257	19270	GCA_001226745.1	1.69574	Contig	21	2	1	1	3	2	1	5
136	GCF_001226825	Campylobacter jejuni	OXC6446	19270	GCA_001226825.1	1.59538	Contig	3029	22	53	4	28	363	3	35
137	GCF_001226845	Campylobacter jejuni	OXC6582	19270	GCA_001226845.1	1.73938	Contig	2030	9	2	4	62	4	5	12
138	GCF_001226945	Campylobacter jejuni	OXC6574	19270	GCA_001226945.1	1.76524	Scaffold	574	7	53	2	10	11	3	3
139	GCF_001227065	Campylobacter jejuni	OXC6573	19270	GCA_001227065.1	1.7047	Contig	21	2	1	1	3	2	1	5
140	GCF_001227125	Campylobacter jejuni	OXC6341	19270	GCA_001227125.1	1.78233	Contig	904	24	2	5	53	23	3	1
141	GCF_001227265	Campylobacter jejuni	OXC6478	19270	GCA_001227265.1	1.68581	Contig	5	7	2	5	2	10	3	6
142	GCF_001227285	Campylobacter jejuni	OXC6625	19270	GCA_001227285.1	1.70334	Contig	21	2	1	1	3	2	1	5
143	GCF_001227305	Campylobacter jejuni	OXC6332	19270	GCA_001227305.1	1.63756	Contig	677	10	81	50	99	120	76	52
144	GCF_001227345	Campylobacter jejuni	OXC6389	19270	GCA_001227345.1	1.64101	Contig	475	2	4	1	4	19	62	5
145	GCF_001227405	Campylobacter jejuni	OXC6516	19270	GCA_001227405.1	1.69535	Contig	53	2	1	21	3	2	1	5
146	GCF_001227425	Campylobacter jejuni	OXC6586	19270	GCA_001227425.1	1.74211	Scaffold	400	8	17	5	2	10	59	6
147	GCF_001227465	Campylobacter jejuni	OXC6304	19270	GCA_001227465.1	1.73874	Contig	574	7	53	2	10	11	3	3
148	GCF_001227485	Campylobacter jejuni	OXC6484	19270	GCA_001227485.1	1.74238	Contig	403	10	27	16	19	10	5	7
149	GCF_001227525	Campylobacter jejuni	OXC6354	19270	GCA_001227525.1	1.64327	Contig	48	2	4	1	2	7	1	5
150	GCF_001227605	Campylobacter jejuni	OXC6611	19270	GCA_001227605.1	1.73437	Contig	353	7	17	5	2	10	3	6
151	GCF_001227665	Campylobacter jejuni	OXC6301	19270	GCA_001227665.1	1.69232	Contig	861	2	1	42	3	148	1	5
152	GCF_001227765	Campylobacter jejuni	OXC6462	19270	GCA_001227765.1	1.69233	Contig	572	62	4	5	2	2	1	5
153	GCF_001227785	Campylobacter jejuni	OXC6379	19270	GCA_001227785.1	1.61766	Contig	5726	2	1	5	462	2	1	5
154	GCF_001227805	Campylobacter jejuni	OXC6454	19270	GCA_001227805.1	1.70418	Scaffold	572	62	4	5	2	2	1	5
155	GCF_001227845	Campylobacter jejuni	OXC6406	19270	GCA_001227845.1	1.63697	Contig	1947	1	94	6	4	3	3	3
156	GCF_001228005	Campylobacter jejuni	OXC6448	19270	GCA_001228005.1	1.6377	Contig	137	4	7	10	4	42	7	1
157	GCF_001228125	Campylobacter jejuni	OXC6353	19270	GCA_001228125.1	1.67004	Scaffold	5729	7	114	2	83	2	3	6
158	GCF_001228165	Campylobacter jejuni	OXC6590	19270	GCA_001228165.1	1.66721	Contig	50	2	1	12	3	2	1	5
159	GCF_001228225	Campylobacter jejuni	OXC6469	19270	GCA_001228225.1	1.73394	Contig	2030	9	2	4	62	4	5	12
160	GCF_001228245	Campylobacter jejuni	OXC6522	19270	GCA_001228245.1	1.67206	Contig	206	2	21	5	37	2	1	5
161	GCF_001228305	Campylobacter jejuni	OXC6256	19270	GCA_001228305.1	1.83197	Scaffold	2274	9	17	5	10	350	3	3
162	GCF_001228425	Campylobacter jejuni	OXC6623	19270	GCA_001228425.1	1.75243	Contig	222	2	21	5	2	59	1	5
163	GCF_001228445	Campylobacter jejuni	OXC6463	19270	GCA_001228445.1	1.61662	Contig	1709	22	15	4	64	74	25	23
164	GCF_001228485	Campylobacter jejuni	OXC6325	19270	GCA_001228485.1	1.69332	Scaffold	19	2	1	5	3	2	1	5
165	GCF_001228525	Campylobacter jejuni	OXC6326	19270	GCA_001228525.1	1.66278	Contig	50	2	1	12	3	2	1	5
166	GCF_001228545	Campylobacter jejuni	OXC6506	19270	GCA_001228545.1	1.64397	Contig	1044	2	10	2	4	19	3	6
167	GCF_001228565	Campylobacter jejuni	OXC6636	19270	GCA_001228565.1	1.66628	Contig	50	2	1	12	3	2	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
168	GCF_001228585	Campylobacter jejuni	OXC6535	19270	GCA_001228585.1	1.64569	Contig	22	1	3	6	4	3	3	3
169	GCF_001228605	Campylobacter jejuni	OXC6420	19270	GCA_001228605.1	1.7072	Scaffold	21	2	1	1	3	2	1	5
170	GCF_001228625	Campylobacter jejuni	OXC6578	19270	GCA_001228625.1	1.63262	Contig	312	14	45	2	4	19	3	6
171	GCF_001228665	Campylobacter jejuni	OXC6383	19270	GCA_001228665.1	1.74069	Contig	21	2	1	1	3	2	1	5
172	GCF_001228745	Campylobacter jejuni	OXC6411	19270	GCA_001228745.1	1.64151	Contig	48	2	4	1	2	7	1	5
173	GCF_001228805	Campylobacter jejuni	OXC6572	19270	GCA_001228805.1	1.78218	Contig	464	24	2	2	2	10	3	1
174	GCF_001228825	Campylobacter jejuni	OXC6526	19270	GCA_001228825.1	1.69251	Contig	990	9	2	4	62	4	133	6
175	GCF_001228865	Campylobacter jejuni	OXC6432	19270	GCA_001228865.1	1.76473	Contig	614	73	21	2	10	86	3	6
176	GCF_001228945	Campylobacter jejuni	OXC6384	19270	GCA_001228945.1	1.61656	Contig	19	2	1	5	3	2	1	5
177	GCF_001229025	Campylobacter jejuni	OXC6275	19270	GCA_001229025.1	1.65826	Contig	21	2	1	1	3	2	1	5
178	GCF_001229065	Campylobacter jejuni	OXC6492	19270	GCA_001229065.1	1.64751	Contig	1044	2	10	2	4	19	3	6
179	GCF_001229085	Campylobacter jejuni	OXC6360	19270	GCA_001229085.1	1.75507	Contig	2274	9	17	5	10	350	3	3
180	GCF_001229145	Campylobacter jejuni	OXC6618	19270	GCA_001229145.1	1.6856	Contig	1900	7	4	2	2	19	1	6
181	GCF_001229225	Campylobacter jejuni	OXC6324	19270	GCA_001229225.1	1.69548	Scaffold	53	2	1	21	3	2	1	5
182	GCF_001229245	Campylobacter jejuni	OXC6615	19270	GCA_001229245.1	1.64026	Contig	50	2	1	12	3	2	1	5
183	GCF_001229325	Campylobacter jejuni	OXC6405	19270	GCA_001229325.1	1.7877	Contig	21	2	1	1	3	2	1	5
184	GCF_001229365	Campylobacter jejuni	OXC6490	19270	GCA_001229365.1	1.7401	Contig	61	1	4	2	2	6	3	17
185	GCF_001229445	Campylobacter jejuni	OXC6473	19270	GCA_001229445.1	1.68725	Contig	5	7	2	5	2	10	3	6
186	GCF_001229465	Campylobacter jejuni	OXC6367	19270	GCA_001229465.1	1.69378	Scaffold	53	2	1	21	3	2	1	5
187	GCF_001229485	Campylobacter jejuni	OXC6362	19270	GCA_001229485.1	1.71113	Contig	257	9	2	4	62	4	5	6
188	GCF_001229505	Campylobacter jejuni	OXC6558	19270	GCA_001229505.1	1.69418	Contig	53	2	1	21	3	2	1	5
189	GCF_001229545	Campylobacter jejuni	OXC6315	19270	GCA_001229545.1	1.71456	Contig	206	2	21	5	37	2	1	5
190	GCF_001229645	Campylobacter jejuni	OXC6449	19270	GCA_001229645.1	1.62957	Contig	50	2	1	12	3	2	1	5
191	GCF_001229665	Campylobacter jejuni	OXC6397	19270	GCA_001229665.1	1.7096	Contig	122	6	4	5	2	2	1	5
192	GCF_001229685	Campylobacter jejuni	OXC6415	19270	GCA_001229685.1	1.65707	Contig	42	1	2	3	4	5	9	3
193	GCF_001229825	Campylobacter jejuni	OXC6281	19270	GCA_001229825.1	1.83115	Contig	2274	9	17	5	10	350	3	3
194	GCF_001229945	Campylobacter jejuni	OXC6515	19270	GCA_001229945.1	1.6336	Contig	845	4	7	73	4	1	7	1
195	GCF_001229985	Campylobacter jejuni	OXC6481	19270	GCA_001229985.1	1.71068	Contig	2361	7	254	2	15	23	3	12
196	GCF_001230045	Campylobacter jejuni	OXC6608	19270	GCA_001230045.1	1.64972	Contig	1073	8	10	2	2	89	12	6
197	GCF_001230085	Campylobacter jejuni	OXC6270	19270	GCA_001230085.1	1.65936	Scaffold	21	2	1	1	3	2	1	5
198	GCF_001230125	Campylobacter jejuni	OXC6429	19270	GCA_001230125.1	1.69661	Contig	5018	2	1	1	3	492	1	5
199	GCF_001230165	Campylobacter jejuni	OXC6621	19270	GCA_001230165.1	1.81223	Contig	464	24	2	2	2	10	3	1
200	GCF_001230205	Campylobacter jejuni	OXC6394	19270	GCA_001230205.1	1.69847	Contig	19	2	1	5	3	2	1	5
201	GCF_001230245	Campylobacter jejuni	OXC6359	19270	GCA_001230245.1	1.87145	Contig	904	24	2	5	53	23	3	1
202	GCF_001230285	Campylobacter jejuni	OXC6534	19270	GCA_001230285.1	1.64484	Contig	122	6	4	5	2	2	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
203	GCF_001230345	Campylobacter jejuni	OXC6466	19270	GCA_001230345.1	1.78644	Scaffold	904	24	2	5	53	23	3	1
204	GCF_001230365	Campylobacter jejuni	OXC6456	19270	GCA_001230365.1	1.73429	Contig	353	7	17	5	2	10	3	6
205	GCF_001230405	Campylobacter jejuni	OXC6313	19270	GCA_001230405.1	1.60874	Contig	45	4	7	10	4	1	7	1
206	GCF_001230525	Campylobacter jejuni	OXC6533	19270	GCA_001230525.1	1.65885	Contig	658	2	4	2	4	19	3	6
207	GCF_001230545	Campylobacter jejuni	OXC6502	19270	GCA_001230545.1	1.65934	Contig	50	2	1	12	3	2	1	5
208	GCF_001230565	Campylobacter jejuni	OXC6303	19270	GCA_001230565.1	1.65365	Contig	19	2	1	5	3	2	1	5
209	GCF_001230585	Campylobacter jejuni	OXC6347	19270	GCA_001230585.1	1.67722	Contig	50	2	1	12	3	2	1	5
210	GCF_001230605	Campylobacter jejuni	OXC6512	19270	GCA_001230605.1	1.58482	Contig	42	1	2	3	4	5	9	3
211	GCF_001230685	Campylobacter jejuni	OXC6539	19270	GCA_001230685.1	1.77397	Contig	21	2	1	1	3	2	1	5
212	GCF_001230705	Campylobacter jejuni	OXC6602	19270	GCA_001230705.1	1.66352	Contig	21	2	1	1	3	2	1	5
213	GCF_001230725	Campylobacter jejuni	OXC6440	19270	GCA_001230725.1	1.71547	Contig	51	7	17	2	15	23	3	12
214	GCF_001230865	Campylobacter jejuni	OXC6483	19270	GCA_001230865.1	1.69996	Scaffold	21	2	1	1	3	2	1	5
215	GCF_001230905	Campylobacter jejuni	OXC6298	19270	GCA_001230905.1	1.74618	Contig	354	8	10	2	2	11	12	6
216	GCF_001230945	Campylobacter jejuni	OXC6638	19270	GCA_001230945.1	1.68434	Contig	48	2	4	1	2	7	1	5
217	GCF_001231005	Campylobacter jejuni	OXC6585	19270	GCA_001231005.1	1.68848	Contig	5	7	2	5	2	10	3	6
218	GCF_001231045	Campylobacter jejuni	OXC6634	19270	GCA_001231045.1	1.69139	Contig	882	7	4	5	68	93	3	46
219	GCF_001231085	Campylobacter jejuni	OXC6422	19270	GCA_001231085.1	1.70301	Contig	572	62	4	5	2	2	1	5
220	GCF_001231105	Campylobacter jejuni	OXC6419	19270	GCA_001231105.1	1.60475	Contig	45	4	7	10	4	1	7	1
221	GCF_001231165	Campylobacter jejuni	OXC6475	19270	GCA_001231165.1	1.71183	Contig	5136	24	2	2	2	10	3	3
222	GCF_001231185	Campylobacter jejuni	OXC6431	19270	GCA_001231185.1	1.6608	Contig	1709	22	15	4	64	74	25	23
223	GCF_001231205	Campylobacter jejuni	OXC6374	19270	GCA_001231205.1	1.7104	Contig	5739	315	112	5	2	13	1	26
224	GCF_001231245	Campylobacter jejuni	OXC6438	19270	GCA_001231245.1	1.72337	Contig	5	7	2	5	2	10	3	6
225	GCF_001231345	Campylobacter jejuni	OXC6370	19270	GCA_001231345.1	1.61786	Contig	19	2	1	5	3	2	1	5
226	GCF_001231385	Campylobacter jejuni	OXC6501	19270	GCA_001231385.1	1.64137	Contig	48	2	4	1	2	7	1	5
227	GCF_001231405	Campylobacter jejuni	OXC6435	19270	GCA_001231405.1	1.83256	Scaffold	573	7	28	4	28	17	34	12
228	GCF_001231425	Campylobacter jejuni	OXC6335	19270	GCA_001231425.1	1.71039	Contig	21	2	1	1	3	2	1	5
229	GCF_001231505	Campylobacter jejuni	OXC6348	19270	GCA_001231505.1	1.65516	Contig	81	2	4	2	2	6	3	17
230	GCF_001231585	Campylobacter jejuni	OXC6401	19270	GCA_001231585.1	1.64337	Contig	257	9	2	4	62	4	5	6
231	GCF_001231605	Campylobacter jejuni	OXC6382	19270	GCA_001231605.1	1.6934	Contig	572	62	4	5	2	2	1	5
232	GCF_001231625	Campylobacter jejuni	OXC6311	19270	GCA_001231625.1	1.70101	Contig	51	7	17	2	15	23	3	12
233	GCF_001231685	Campylobacter jejuni	OXC6450	19270	GCA_001231685.1	1.68176	Contig	257	9	2	4	62	4	5	6
234	GCF_001231705	Campylobacter jejuni	OXC6632	19270	GCA_001231705.1	1.68935	Contig	61	1	4	2	2	6	3	17
235	GCF_001231745	Campylobacter jejuni	OXC6553	19270	GCA_001231745.1	1.75115	Contig	273	2	21	5	37	60	1	5
236	GCF_001231765	Campylobacter jejuni	OXC6468	19270	GCA_001231765.1	1.74371	Contig	5732	24	2	2	2	10	59	1
237	GCF_001231785	Campylobacter jejuni	OXC6284	19270	GCA_001231785.1	1.68605	Contig	48	2	4	1	2	7	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
238	GCF_001231845	Campylobacter jejuni	OXC6575	19270	GCA_001231845.1	1.72987	Contig	5	7	2	5	2	10	3	6
239	GCF_001231905	Campylobacter jejuni	OXC6464	19270	GCA_001231905.1	1.66054	Contig	5018	2	1	1	3	492	1	5
240	GCF_001231985	Campylobacter jejuni	OXC6331	19270	GCA_001231985.1	1.67688	Contig	50	2	1	12	3	2	1	5
241	GCF_001232125	Campylobacter jejuni	OXC6299	19270	GCA_001232125.1	1.74632	Contig	354	8	10	2	2	11	12	6
242	GCF_001232145	Campylobacter jejuni	OXC6567	19270	GCA_001232145.1	1.74511	Contig	354	8	10	2	2	11	12	6
243	GCF_001232165	Campylobacter jejuni	OXC6641	19270	GCA_001232165.1	1.61966	Contig	22	1	3	6	4	3	3	3
244	GCF_001232205	Campylobacter jejuni	OXC6549	19270	GCA_001232205.1	1.64598	Contig	48	2	4	1	2	7	1	5
245	GCF_001232265	Campylobacter jejuni	OXC6293	19270	GCA_001232265.1	1.63785	Contig	45	4	7	10	4	1	7	1
246	GCF_001232365	Campylobacter jejuni	OXC6560	19270	GCA_001232365.1	1.75969	Contig	904	24	2	5	53	23	3	1
247	GCF_001232405	Campylobacter jejuni	OXC6250	19270	GCA_001232405.1	1.68654	Contig	257	9	2	4	62	4	5	6
248	GCF_001232425	Campylobacter jejuni	OXC6409	19270	GCA_001232425.1	1.58514	Contig	42	1	2	3	4	5	9	3
249	GCF_001232465	Campylobacter jejuni	OXC6352	19270	GCA_001232465.1	1.78131	Contig	464	24	2	2	2	10	3	1
250	GCF_001232485	Campylobacter jejuni	OXC6342	19270	GCA_001232485.1	1.66482	Contig	51	7	17	2	15	23	3	12
251	GCF_001232545	Campylobacter jejuni	OXC6494	19270	GCA_001232545.1	1.7333	Contig	353	7	17	5	2	10	3	6
252	GCF_001232585	Campylobacter jejuni	OXC6402	19270	GCA_001232585.1	1.64159	Contig	61	1	4	2	2	6	3	17
253	GCF_001232665	Campylobacter jejuni	OXC6550	19270	GCA_001232665.1	1.84052	Scaffold	464	24	2	2	2	10	3	1
254	GCF_001232705	Campylobacter jejuni	OXC6365	19270	GCA_001232705.1	1.60599	Contig	45	4	7	10	4	1	7	1
255	GCF_001232725	Campylobacter jejuni	OXC6373	19270	GCA_001232725.1	1.63178	Contig	267	4	7	40	4	42	51	1
256	GCF_001232745	Campylobacter jejuni	OXC6530	19270	GCA_001232745.1	1.66604	Contig	50	2	1	12	3	2	1	5
257	GCF_001232825	Campylobacter jejuni	OXC6563	19270	GCA_001232825.1	1.70355	Contig	21	2	1	1	3	2	1	5
258	GCF_001232885	Campylobacter jejuni	OXC6493	19270	GCA_001232885.1	1.66447	Scaffold	50	2	1	12	3	2	1	5
259	GCF_001232905	Campylobacter jejuni	OXC6461	19270	GCA_001232905.1	1.66564	Scaffold	50	2	1	12	3	2	1	5
260	GCF_001232925	Campylobacter jejuni	OXC6269	19270	GCA_001232925.1	1.71785	Contig	48	2	4	1	2	7	1	5
261	GCF_001232945	Campylobacter jejuni	OXC6491	19270	GCA_001232945.1	1.7245	Contig	257	9	2	4	62	4	5	6
262	GCF_001232965	Campylobacter jejuni	OXC6629	19270	GCA_001232965.1	1.70322	Contig	21	2	1	1	3	2	1	5
263	GCF_001233005	Campylobacter jejuni	OXC6279	19270	GCA_001233005.1	1.64427	Contig	2258	2	84	2	68	2	68	5
264	GCF_001233025	Campylobacter jejuni	OXC6603	19270	GCA_001233025.1	1.69707	Contig	19	2	1	5	3	2	1	5
265	GCF_001233065	Campylobacter jejuni	OXC6503	19270	GCA_001233065.1	1.66899	Contig	51	7	17	2	15	23	3	12
266	GCF_001233085	Campylobacter jejuni	OXC6500	19270	GCA_001233085.1	1.72033	Contig	2135	8	1	6	3	2	1	12
267	GCF_001233105	Campylobacter jejuni	OXC6592	19270	GCA_001233105.1	1.64979	Contig	137	4	7	10	4	42	7	1
268	GCF_001233185	Campylobacter jejuni	OXC6412	19270	GCA_001233185.1	1.69336	Contig	257	9	2	4	62	4	5	6
269	GCF_001233225	Campylobacter jejuni	OXC6528	19270	GCA_001233225.1	1.62349	Contig	5738	160	84	5	10	13	3	5
270	GCF_001233245	Campylobacter jejuni	OXC6290	19270	GCA_001233245.1	1.69586	Contig	257	9	2	4	62	4	5	6
271	GCF_001233285	Campylobacter jejuni	OXC6272	19270	GCA_001233285.1	1.68947	Contig	844	2	4	1	3	7	1	5
272	GCF_001233305	Campylobacter jejuni	OXC6381	19270	GCA_001233305.1	1.8668	Contig	523	2	4	1	93	11	3	6

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
273	GCF_001233325	Campylobacter jejuni	OXC6418	19270	GCA_001233325.1	1.69744	Contig	354	8	10	2	2	11	12	6
274	GCF_001233405	Campylobacter jejuni	OXC6570	19270	GCA_001233405.1	1.65143	Contig	61	1	4	2	2	6	3	17
275	GCF_001233465	Campylobacter jejuni	OXC6571	19270	GCA_001233465.1	1.67684	Contig	50	2	1	12	3	2	1	5
276	GCF_001233545	Campylobacter jejuni	OXC6569	19270	GCA_001233545.1	1.63528	Contig	5597	10	2	42	4	90	25	8
277	GCF_001233585	Campylobacter jejuni	OXC6330	19270	GCA_001233585.1	1.60754	Contig	583	4	7	10	4	42	51	1
278	GCF_001233625	Campylobacter jejuni	OXC6264	19270	GCA_001233625.1	1.68117	Contig	22	1	3	6	4	3	3	3
279	GCF_001233645	Campylobacter jejuni	OXC6350	19270	GCA_001233645.1	1.62752	Contig	42	1	2	3	4	5	9	3
280	GCF_001233665	Campylobacter jejuni	OXC6390	19270	GCA_001233665.1	1.6815	Contig	257	9	2	4	62	4	5	6
281	GCF_001233685	Campylobacter jejuni	OXC6531	19270	GCA_001233685.1	1.66747	Scaffold	50	2	1	12	3	2	1	5
282	GCF_001233745	Campylobacter jejuni	OXC6604	19270	GCA_001233745.1	1.70183	Scaffold	21	2	1	1	3	2	1	5
283	GCF_001233765	Campylobacter jejuni	OXC6314	19270	GCA_001233765.1	1.60977	Scaffold	45	4	7	10	4	1	7	1
284	GCF_001233805	Campylobacter jejuni	OXC6584	19270	GCA_001233805.1	1.81323	Contig	5758	7	422	5	10	11	3	6
285	GCF_001233905	Campylobacter jejuni	OXC6294	19270	GCA_001233905.1	1.6083	Contig	583	4	7	10	4	42	51	1
286	GCF_001233965	Campylobacter jejuni	OXC6251	19270	GCA_001233965.1	1.6577	Contig	21	2	1	1	3	2	1	5
287	GCF_001233985	Campylobacter jejuni	OXC6599	19270	GCA_001233985.1	1.62398	Contig	2314	2	61	4	64	332	7	23
288	GCF_001234065	Campylobacter jejuni	OXC6288	19270	GCA_001234065.1	1.73601	Contig	464	24	2	2	2	10	3	1
289	GCF_001234085	Campylobacter jejuni	OXC6339	19270	GCA_001234085.1	1.61023	Contig	45	4	7	10	4	1	7	1
290	GCF_001234145	Campylobacter jejuni	OXC6292	19270	GCA_001234145.1	1.67548	Contig	50	2	1	12	3	2	1	5
291	GCF_001234165	Campylobacter jejuni	OXC6612	19270	GCA_001234165.1	1.73152	Contig	257	9	2	4	62	4	5	6
292	GCF_001234225	Campylobacter jejuni	OXC6285	19270	GCA_001234225.1	1.65875	Contig	21	2	1	1	3	2	1	5
293	GCF_001234285	Campylobacter jejuni	OXC6565	19270	GCA_001234285.1	1.66282	Contig	50	2	1	12	3	2	1	5
294	GCF_001234305	Campylobacter jejuni	OXC6322	19270	GCA_001234305.1	1.8011	Contig	2274	9	17	5	10	350	3	3
295	GCF_001234325	Campylobacter jejuni	OXC6561	19270	GCA_001234325.1	1.68592	Contig	132	1	6	22	24	12	28	1
296	GCF_001234425	Campylobacter jejuni	OXC6639	19270	GCA_001234425.1	1.64942	Contig	132	1	6	22	24	12	28	1
297	GCF_001234505	Campylobacter jejuni	OXC6261	19270	GCA_001234505.1	1.61936	Scaffold	5717	2	421	4	64	332	7	23
298	GCF_001234565	Campylobacter jejuni	OXC6316	19270	GCA_001234565.1	1.70867	Scaffold	51	7	17	2	15	23	3	12
299	GCF_001234605	Campylobacter jejuni	OXC6467	19270	GCA_001234605.1	1.73359	Contig	2030	9	2	4	62	4	5	12
300	GCF_001234665	Campylobacter jejuni	OXC6283	19270	GCA_001234665.1	1.65116	Contig	61	1	4	2	2	6	3	17
301	GCF_001234685	Campylobacter jejuni	OXC6413	19270	GCA_001234685.1	1.66549	Scaffold	429	7	4	5	2	11	1	5
302	GCF_001234745	Campylobacter jejuni	OXC6427	19270	GCA_001234745.1	1.76327	Contig	2844	24	21	2	2	2	59	6
303	GCF_001234825	Campylobacter jejuni	OXC6482	19270	GCA_001234825.1	1.68409	Contig	220	1	6	29	2	40	32	3
304	GCF_001234865	Campylobacter jejuni	OXC6268	19270	GCA_001234865.1	1.69008	Contig	61	1	4	2	2	6	3	17
305	GCF_001234885	Campylobacter jejuni	OXC6407	19270	GCA_001234885.1	1.60632	Contig	257	9	2	4	62	4	5	6
306	GCF_001234945	Campylobacter jejuni	OXC6541	19270	GCA_001234945.1	1.7188	Contig	5136	24	2	2	2	10	3	3
307	GCF_001235005	Campylobacter jejuni	OXC6627	19270	GCA_001235005.1	1.70511	Contig	21	2	1	1	3	2	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
308	GCF_001235045	Campylobacter jejuni	OXC6556	19270	GCA_001235045.1	1.69212	Contig	4811	6	4	52	2	89	282	5
309	GCF_001235065	Campylobacter jejuni	OXC6291	19270	GCA_001235065.1	1.73287	Contig	2030	9	2	4	62	4	5	12
310	GCF_001235125	Campylobacter jejuni	OXC6505	19270	GCA_001235125.1	1.6741	Contig	141	2	1	10	3	2	1	5
311	GCF_001235145	Campylobacter jejuni	OXC6318	19270	GCA_001235145.1	1.71581	Contig	356	14	17	5	2	11	3	6
312	GCF_001235165	Campylobacter jejuni	OXC6423	19270	GCA_001235165.1	1.70929	Contig	464	24	2	2	2	10	3	1
313	GCF_001235245	Campylobacter jejuni	OXC6465	19270	GCA_001235245.1	1.70117	Contig	572	62	4	5	2	2	1	5
314	GCF_001235365	Campylobacter jejuni	OXC6540	19270	GCA_001235365.1	1.63116	Contig	587	1	2	42	4	90	25	8
315	GCF_001235565	Campylobacter jejuni	OXC6510	19270	GCA_001235565.1	1.5937	Contig	267	4	7	40	4	42	51	1
316	GCF_001235585	Campylobacter jejuni	OXC6547	19270	GCA_001235585.1	1.64211	Contig	48	2	4	1	2	7	1	5
317	GCF_001235625	Campylobacter jejuni	OXC6480	19270	GCA_001235625.1	1.57072	Contig	1301	2	115	57	26	127	29	35
318	GCF_001235765	Campylobacter jejuni	OXC6410	19270	GCA_001235765.1	1.75188	Contig	5728	2	59	4	38	10	488	35
319	GCF_001235825	Campylobacter jejuni	OXC6344	19270	GCA_001235825.1	1.68116	Contig	257	9	2	4	62	4	5	6
320	GCF_001235925	Campylobacter jejuni	OXC6499	19270	GCA_001235925.1	1.73352	Contig	2030	9	2	4	62	4	5	12
321	GCF_001235965	Campylobacter jejuni	OXC6317	19270	GCA_001235965.1	1.828	Contig	21	2	1	1	3	2	1	5
322	GCF_001235985	Campylobacter jejuni	OXC6548	19270	GCA_001235985.1	1.65944	Contig	19	2	1	5	3	2	1	5
323	GCF_001236085	Campylobacter jejuni	OXC6357	19270	GCA_001236085.1	1.63899	Contig	51	7	17	2	15	23	3	12
324	GCF_001236105	Campylobacter jejuni	OXC6255	19270	GCA_001236105.1	1.70026	Contig	21	2	1	1	3	2	1	5
325	GCF_001236125	Campylobacter jejuni	OXC6557	19270	GCA_001236125.1	1.7286	Contig	5756	73	21	2	465	86	3	6
326	GCF_001236165	Campylobacter jejuni	OXC6613	19270	GCA_001236165.1	1.72602	Contig	50	2	1	12	3	2	1	5
327	GCF_001236205	Campylobacter jejuni	OXC6544	19270	GCA_001236205.1	1.73123	Contig	1911	7	84	5	10	119	178	26
328	GCF_001236225	Campylobacter jejuni	OXC6457	19270	GCA_001236225.1	1.73002	Scaffold	50	2	1	12	3	2	1	5
329	GCF_001236305	Campylobacter jejuni	OXC6517	19270	GCA_001236305.1	1.70636	Contig	354	8	10	2	2	11	12	6
330	GCF_001236325	Campylobacter jejuni	OXC6273	19270	GCA_001236325.1	1.65465	Contig	508	1	6	60	24	12	28	1
331	GCF_001236345	Campylobacter jejuni	OXC6524	19270	GCA_001236345.1	1.66624	Contig	50	2	1	12	3	2	1	5
332	GCF_001236385	Campylobacter jejuni	OXC6282	19270	GCA_001236385.1	1.71996	Contig	53	2	1	21	3	2	1	5
333	GCF_001236445	Campylobacter jejuni	OXC6351	19270	GCA_001236445.1	1.64471	Contig	45	4	7	10	4	1	7	1
334	GCF_001236525	Campylobacter jejuni	OXC6433	19270	GCA_001236525.1	1.83456	Scaffold	573	7	28	4	28	17	34	12
335	GCF_001236545	Campylobacter jejuni	OXC6562	19270	GCA_001236545.1	1.74202	Contig	21	2	1	1	3	2	1	5
336	GCF_001236605	Campylobacter jejuni	OXC6392	19270	GCA_001236605.1	1.76817	Contig	574	7	53	2	10	11	3	3
337	GCF_001236645	Campylobacter jejuni	OXC6581	19270	GCA_001236645.1	1.78608	Contig	464	24	2	2	2	10	3	1
338	GCF_001236665	Campylobacter jejuni	OXC6518	19270	GCA_001236665.1	1.77343	Contig	354	8	10	2	2	11	12	6
339	GCF_001236705	Campylobacter jejuni	OXC6368	19270	GCA_001236705.1	1.72278	Contig	52	9	25	2	10	22	3	6
340	GCF_001236725	Campylobacter jejuni	OXC6631	19270	GCA_001236725.1	1.75872	Scaffold	2401	64	105	20	287	94	103	16
341	GCF_001236765	Campylobacter jejuni	OXC6488	19270	GCA_001236765.1	1.64729	Contig	48	2	4	1	2	7	1	5
342	GCF_001236785	Campylobacter jejuni	OXC6396	19270	GCA_001236785.1	1.67764	Contig	51	7	17	2	15	23	3	12

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
343	GCF_001236805	Campylobacter jejuni	OXC6511	19270	GCA_001236805.1	1.59377	Contig	267	4	7	40	4	42	51	1
344	GCF_001236845	Campylobacter jejuni	OXC6487	19270	GCA_001236845.1	1.64926	Contig	48	2	4	1	2	7	1	5
345	GCF_001236865	Campylobacter jejuni	OXC6259	19270	GCA_001236865.1	1.64593	Contig	45	4	7	10	4	1	7	1
346	GCF_001236905	Campylobacter jejuni	OXC6600	19270	GCA_001236905.1	1.66159	Contig	50	2	1	12	3	2	1	5
347	GCF_001236945	Campylobacter jejuni	OXC6437	19270	GCA_001236945.1	1.64309	Contig	45	4	7	10	4	1	7	1
348	GCF_001237005	Campylobacter jejuni	OXC6323	19270	GCA_001237005.1	1.73947	Contig	574	7	53	2	10	11	3	3
349	GCF_001237085	Campylobacter jejuni	OXC6340	19270	GCA_001237085.1	1.73084	Contig	262	2	1	1	3	2	1	3
350	GCF_001237125	Campylobacter jejuni	OXC6635	19270	GCA_001237125.1	1.71669	Scaffold	5136	24	2	2	2	10	3	3
351	GCF_001237145	Campylobacter jejuni	OXC6403	19270	GCA_001237145.1	1.73241	Contig	48	2	4	1	2	7	1	5
352	GCF_001237205	Campylobacter jejuni	OXC6364	19270	GCA_001237205.1	1.65683	Contig	132	1	6	22	24	12	28	1
353	GCF_001237265	Campylobacter jejuni	OXC6277	19270	GCA_001237265.1	1.65715	Contig	50	2	1	12	3	2	1	5
354	GCF_001237345	Campylobacter jejuni	OXC6458	19270	GCA_001237345.1	1.76205	Scaffold	5731	24	2	2	2	10	3	12
355	GCF_001237385	Campylobacter jejuni	OXC6470	19270	GCA_001237385.1	1.67587	Contig	572	62	4	5	2	2	1	5
356	GCF_001237465	Campylobacter jejuni	OXC6591	19270	GCA_001237465.1	1.72437	Contig	5	7	2	5	2	10	3	6
357	GCF_001237545	Campylobacter jejuni	OXC6361	19270	GCA_001237545.1	1.72881	Contig	574	7	53	2	10	11	3	3
358	GCF_001237565	Campylobacter jejuni	OXC6616	19270	GCA_001237565.1	1.74061	Contig	21	2	1	1	3	2	1	5
359	GCF_001237585	Campylobacter jejuni	OXC6441	19270	GCA_001237585.1	1.67434	Contig	52	9	25	2	10	22	3	6
360	GCF_001237605	Campylobacter jejuni	OXC6375	19270	GCA_001237605.1	1.73112	Contig	257	9	2	4	62	4	5	6
361	GCF_001237645	Campylobacter jejuni	OXC6363	19270	GCA_001237645.1	1.61222	Contig	658	2	4	2	4	19	3	6
362	GCF_001237745	Campylobacter jejuni	OXC6451	19270	GCA_001237745.1	1.73826	Contig	354	8	10	2	2	11	12	6
363	GCF_001237785	Campylobacter jejuni	OXC6321	19270	GCA_001237785.1	1.65432	Scaffold	45	4	7	10	4	1	7	1
364	GCF_001237865	Campylobacter jejuni	OXC6489	19270	GCA_001237865.1	1.62728	Scaffold	50	2	1	12	3	2	1	5
365	GCF_001237885	Campylobacter jejuni	OXC6485	19270	GCA_001237885.1	1.74486	Contig	403	10	27	16	19	10	5	7
366	GCF_001237965	Campylobacter jejuni	OXC6495	19270	GCA_001237965.1	1.71019	Contig	572	62	4	5	2	2	1	5
367	GCF_001238065	Campylobacter jejuni	OXC6286	19270	GCA_001238065.1	1.6761	Contig	50	2	1	12	3	2	1	5
368	GCF_001238105	Campylobacter jejuni	OXC6486	19270	GCA_001238105.1	1.79899	Scaffold	2274	9	17	5	10	350	3	3
369	GCF_001238125	Campylobacter jejuni	OXC6624	19270	GCA_001238125.1	1.61239	Contig	45	4	7	10	4	1	7	1
370	GCF_001238165	Campylobacter jejuni	OXC6479	19270	GCA_001238165.1	1.6643	Contig	883	2	17	2	3	2	1	5
371	GCF_001238185	Campylobacter jejuni	OXC6619	19270	GCA_001238185.1	1.69529	Contig	21	2	1	1	3	2	1	5
372	GCF_001238245	Campylobacter jejuni	OXC6310	19270	GCA_001238245.1	1.65816	Contig	21	2	1	1	3	2	1	5
373	GCF_001238565	Campylobacter jejuni	OXC6404	19270	GCA_001238565.1	1.73402	Contig	2030	9	2	4	62	4	5	12
374	GCF_001238765	Campylobacter jejuni	OXC6514	19270	GCA_001238765.1	1.65222	Contig	3102	2	84	12	3	11	1	5
375	GCF_001238805	Campylobacter jejuni	OXC6388	19270	GCA_001238805.1	1.62795	Contig	42	1	2	3	4	5	9	3
376	GCF_001238825	Campylobacter jejuni	OXC6369	19270	GCA_001238825.1	1.59429	Contig	267	4	7	40	4	42	51	1
377	GCF_001238885	Campylobacter jejuni	OXC6637	19270	GCA_001238885.1	1.74831	Scaffold	21	2	1	1	3	2	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
378	GCF_001238945	Campylobacter jejuni	OXC6349	19270	GCA_001238945.1	1.82358	Contig	933	10	1	59	19	10	5	7
379	GCF_001238965	Campylobacter jejuni	OXC6507	19270	GCA_001238965.1	1.73018	Scaffold	5	7	2	5	2	10	3	6
380	GCF_001238985	Campylobacter jejuni	OXC6610	19270	GCA_001238985.1	1.83909	Scaffold	573	7	28	4	28	17	34	12
381	GCF_001239025	Campylobacter jejuni	OXC6393	19270	GCA_001239025.1	1.66497	Contig	5727	2	1	12	462	2	1	5
382	GCF_001239045	Campylobacter jejuni	OXC6546	19270	GCA_001239045.1	1.78174	Scaffold	449	7	71	5	62	11	67	6
383	GCF_001239105	Campylobacter jejuni	OXC6477	19270	GCA_001239105.1	1.70364	Scaffold	572	62	4	5	2	2	1	5
384	GCF_001239205	Campylobacter jejuni	OXC6555	19270	GCA_001239205.1	1.84275	Scaffold	464	24	2	2	2	10	3	1
385	GCF_001239265	Campylobacter jejuni	OXC6595	19270	GCA_001239265.1	1.83249	Contig	573	7	28	4	28	17	34	12
386	GCF_001239305	Campylobacter jejuni	OXC6459	19270	GCA_001239305.1	1.66609	Contig	50	2	1	12	3	2	1	5
387	GCF_001239325	Campylobacter jejuni	OXC6271	19270	GCA_001239325.1	1.65224	Contig	508	1	6	60	24	12	28	1
388	GCF_001239345	Campylobacter jejuni	OXC6525	19270	GCA_001239345.1	1.73038	Scaffold	122	6	4	5	2	2	1	5
389	GCF_001239365	Campylobacter jejuni	OXC6580	19270	GCA_001239365.1	1.70151	Contig	572	62	4	5	2	2	1	5
390	GCF_001239445	Campylobacter jejuni	OXC6274	19270	GCA_001239445.1	1.64131	Scaffold	51	7	17	2	15	23	3	12
391	GCF_001239485	Campylobacter jejuni	OXC6327	19270	GCA_001239485.1	1.73484	Contig	3534	64	81	111	143	134	153	16
392	GCF_001239645	Campylobacter jejuni	OXC6455	19270	GCA_001239645.1	1.82489	Contig	5730	7	17	5	2	22	5	1
393	GCF_001239685	Campylobacter jejuni	OXC6336	19270	GCA_001239685.1	1.72721	Contig	986	91	2	42	4	169	9	8
394	GCF_001291505	Campylobacter jejuni	RC427	19270	GCA_001291505.1	1.76469	Scaffold	814	2	75	4	48	141	34	1
395	GCF_001291565	Campylobacter jejuni	RC188	51038	GCA_001291565.1	1.8017	Scaffold	4425	33	39	30	139	113	47	17
396	GCF_001291585	Campylobacter jejuni	RC507	19270	GCA_001291585.1	1.60139	Scaffold	52	9	25	2	10	22	3	6
397	GCF_001291625	Campylobacter jejuni	RC186	51038	GCA_001291625.1	1.79986	Scaffold	4425	33	39	30	139	113	47	17
398	GCF_001291645	Campylobacter jejuni	RC039	51038	GCA_001291645.1	1.79885	Scaffold	4425	33	39	30	139	113	47	17
399	GCF_001291685	Campylobacter jejuni	RC179	51038	GCA_001291685.1	1.80424	Scaffold	4425	33	39	30	139	113	47	17
400	GCF_001291705	Campylobacter jejuni	RC185	51038	GCA_001291705.1	1.76122	Scaffold	4425	33	39	30	139	113	47	17
401	GCF_001291805	Campylobacter jejuni	RC270	19270	GCA_001291805.1	1.70908	Scaffold	814	2	75	4	48	141	34	1
402	GCF_001291825	Campylobacter jejuni	RC009	19270	GCA_001291825.1	1.6522	Scaffold	132	1	6	22	24	12	28	1
403	GCF_001291925	Campylobacter jejuni	RC429	19270	GCA_001291925.1	1.6244	Scaffold	45	4	7	10	4	1	7	1
404	GCF_001291945	Campylobacter jejuni	RC508	51038	GCA_001291945.1	1.80496	Scaffold	4425	33	39	30	139	113	47	17
405	GCF_001291965	Campylobacter jejuni	12502	19270	GCA_001291965.1	1.68235	Scaffold	5	7	2	5	2	10	3	6
406	GCF_001292025	Campylobacter jejuni	RC169	19270	GCA_001292025.1	1.74971	Scaffold	814	2	75	4	48	141	34	1
407	GCF_001292045	Campylobacter jejuni	RC526	19270	GCA_001292045.1	1.64932	Scaffold	814	2	75	4	48	141	34	1
408	GCF_001292145	Campylobacter jejuni	RC317	19270	GCA_001292145.1	1.80307	Scaffold	814	2	75	4	48	141	34	1
409	GCF_001292185	Campylobacter jejuni	RC517	19270	GCA_001292185.1	1.70161	Scaffold	814	2	75	4	48	141	34	1
410	GCF_001292245	Campylobacter jejuni	RC104	19270	GCA_001292245.1	1.70189	Scaffold	4279	27	22	22	376	43	86	31
411	GCF_001292285	Campylobacter jejuni	RC168	19270	GCA_001292285.1	1.72438	Scaffold	814	2	75	4	48	141	34	1
412	GCF_001292385	Campylobacter jejuni	RC280	19270	GCA_001292385.1	1.79858	Scaffold	814	2	75	4	48	141	34	1

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
413	GCF_001292465	Campylobacter jejuni	RC220	51038	GCA_001292465.1	1.79093	Scaffold	4425	33	39	30	139	113	47	17
414	GCF_001299565	Campylobacter jejuni subsp. jejuni	RM3197	19270	GCA_001299565.1	1.66457	Complete Genome	362	1	2	49	4	11	66	8
415	GCF_001299595	Campylobacter jejuni subsp. jejuni	RM3196	19270	GCA_001299595.1	1.66457	Complete Genome	362	1	2	49	4	11	66	8
416	GCF_001314285	Campylobacter jejuni	RM1285	19270	GCA_001314285.1	1.6358	Complete Genome	50	2	1	12	3	2	1	5
417	GCF_001406895	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001406895.1	1.63551	Scaffold	51	7	17	2	15	23	3	12
418	GCF_001406935	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001406935.1	1.66967	Scaffold	52	9	25	2	10	22	3	6
419	GCF_001406955	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001406955.1	1.6212	Scaffold	22	1	3	6	4	3	3	3
420	GCF_001406975	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001406975.1	1.65913	Scaffold	148	2	1	6	3	2	1	5
421	GCF_001406995	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001406995.1	1.73003	Scaffold	904	24	2	5	53	23	3	1
422	GCF_001407015	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407015.1	1.65771	Scaffold	47	2	1	1	5	2	1	5
423	GCF_001407035	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407035.1	1.65442	Scaffold	51	7	17	2	15	23	3	12
424	GCF_001407055	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407055.1	1.57836	Scaffold	45	4	7	10	4	1	7	1
425	GCF_001407075	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407075.1	1.78155	Scaffold	356	14	17	5	2	11	3	6
426	GCF_001407095	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407095.1	1.7418	Scaffold	464	24	2	2	2	10	3	1
427	GCF_001407115	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407115.1	1.82039	Scaffold	824	9	2	2	2	11	5	6
428	GCF_001407135	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407135.1	1.69384	Scaffold	48	2	4	1	2	7	1	5
429	GCF_001407195	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407195.1	1.66957	Scaffold	122	6	4	5	2	2	1	5
430	GCF_001407315	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407315.1	1.65722	Scaffold	50	2	1	12	3	2	1	5
431	GCF_001407345	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407345.1	1.71916	Scaffold	51	7	17	2	15	23	3	12
432	GCF_001407365	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407365.1	1.66675	Scaffold	21	2	1	1	3	2	1	5
433	GCF_001407415	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407415.1	1.70372	Scaffold	572	62	4	5	2	2	1	5
434	GCF_001407435	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407435.1	1.69516	Scaffold	148	2	1	6	3	2	1	5
435	GCF_001407455	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407455.1	1.7034	Scaffold	464	24	2	2	2	10	3	1
436	GCF_001407475	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407475.1	1.83586	Scaffold	1707	9	2	5	2	11	3	1
437	GCF_001407495	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407495.1	1.72988	Scaffold	45	4	7	10	4	1	7	1
438	GCF_001407515	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407515.1	1.66377	Scaffold	47	2	1	1	5	2	1	5
439	GCF_001407535	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407535.1	1.63421	Scaffold	441	7	1	2	83	2	3	6
440	GCF_001407555	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407555.1	1.6853	Scaffold	122	6	4	5	2	2	1	5
441	GCF_001407575	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407575.1	1.587	Scaffold	42	1	2	3	4	5	9	3
442	GCF_001407595	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407595.1	1.64726	Scaffold	122	6	4	5	2	2	1	5
443	GCF_001407615	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407615.1	1.79771	Scaffold	46	2	21	5	3	2	1	5
444	GCF_001407635	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407635.1	1.68435	Scaffold	354	8	10	2	2	11	12	6
445	GCF_001407655	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407655.1	1.64556	Scaffold	48	2	4	1	2	7	1	5

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
446	GCF_001407675	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407675.1	1.65802	Scaffold	45	4	7	10	4	1	7	1
447	GCF_001407695	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407695.1	1.6466	Scaffold	2180	2	4	2	25	11	3	5
448	GCF_001407715	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407715.1	1.60634	Scaffold	61	1	4	2	2	6	3	17
449	GCF_001407735	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407735.1	1.66658	Scaffold	47	2	1	1	5	2	1	5
450	GCF_001407755	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407755.1	1.71369	Scaffold	572	62	4	5	2	2	1	5
451	GCF_001407775	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407775.1	1.71375	Scaffold	443	24	17	2	15	23	3	12
452	GCF_001407795	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407795.1	1.71094	Scaffold	356	14	17	5	2	11	3	6
453	GCF_001407815	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407815.1	1.7627	Scaffold	904	24	2	5	53	23	3	1
454	GCF_001407855	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407855.1	1.6543	Scaffold	262	2	1	1	3	2	1	3
455	GCF_001407895	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407895.1	1.64074	Scaffold	2123	1	2	42	4	2	25	8
456	GCF_001407915	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407915.1	1.61217	Scaffold	583	4	7	10	4	42	51	1
457	GCF_001407935	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407935.1	1.65527	Scaffold	48	2	4	1	2	7	1	5
458	GCF_001407955	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407955.1	1.67112	Scaffold	883	2	17	2	3	2	1	5
459	GCF_001407975	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407975.1	1.61118	Scaffold	45	4	7	10	4	1	7	1
460	GCF_001407995	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001407995.1	1.58754	Scaffold	42	1	2	3	4	5	9	3
461	GCF_001408015	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408015.1	1.6569	Scaffold	19	2	1	5	3	2	1	5
462	GCF_001408035	Campylobacter jejuni	Campylobacter jejuni	51038	GCA_001408035.1	1.76976	Scaffold	860	33	39	30	79	113	47	17
463	GCF_001408055	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408055.1	1.77208	Scaffold	7546	2	257	80	243	385	25	212
464	GCF_001408095	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408095.1	1.63617	Scaffold	441	7	1	2	83	2	3	6
465	GCF_001408115	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408115.1	1.75419	Scaffold	2364	14	249	5	2	11	3	6
466	GCF_001408135	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408135.1	1.67606	Scaffold	429	7	4	5	2	11	1	5
467	GCF_001408155	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408155.1	1.64835	Scaffold	367	2	2	4	62	4	5	6
468	GCF_001408175	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408175.1	1.65478	Scaffold	52	9	25	2	10	22	3	6
469	GCF_001408195	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408195.1	1.68458	Scaffold	904	24	2	5	53	23	3	1
470	GCF_001408215	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408215.1	1.60772	Scaffold	45	4	7	10	4	1	7	1
471	GCF_001408235	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_001408235.1	1.6147	Scaffold	50	2	1	12	3	2	1	5
472	GCF_001412295	Campylobacter jejuni	CJM1cam	19270	GCA_001412295.1	1.61666	Complete Genome	137	4	7	10	4	42	7	1
473	GCF_001416835	Campylobacter jejuni	CVM N51684	19270	GCA_001416835.1	1.77097	Contig	7729	7	112	5	2	741	67	6
474	GCF_001417165	Campylobacter jejuni	CVM N42547	19270	GCA_001417165.1	1.73027	Contig	222	2	21	5	2	59	1	5
475	GCF_001417285	Campylobacter jejuni	CVM N51691	19270	GCA_001417285.1	1.83931	Contig	2934	7	112	5	2	167	67	6
476	GCF_001418835	Campylobacter jejuni CVM 41900	CVM 41900	19270	GCA_001418835.1	1.68237	Contig	2140	9	53	2	53	11	3	3
477	GCF_001418855	Campylobacter jejuni CVM 41902	CVM 41902	19270	GCA_001418855.1	1.64857	Scaffold	8171	14	21	2	10	127	3	6
478	GCF_001418905	Campylobacter jejuni CVM 41905	CVM 41905	19270	GCA_001418905.1	1.52549	Contig	2109	4	7	10	4	10	7	1
479	GCF_001418915	Campylobacter jejuni CVM 41908	CVM 41908	19270	GCA_001418915.1	1.61622	Contig	2109	4	7	10	4	10	7	1
480	GCF_001418925	Campylobacter jejuni CVM 41910	CVM 41910	19270	GCA_001418925.1	1.71842	Contig	2109	4	7	10	4	10	7	1

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
481	GCF_001418935	Campylobacter jejuni CVM 41912	CVM 41912	19270	GCA_001418935.1	1.7113	Contig	50	2	1	12	3	2	1	5
482	GCF_001419065	Campylobacter jejuni CVM 41921	CVM 41921	19270	GCA_001419065.1	1.74222	Contig	1233	7	17	5	10	10	177	6
483	GCF_001419085	Campylobacter jejuni CVM 41922	CVM 41922	19270	GCA_001419085.1	1.72987	Contig	48	2	4	1	2	7	1	5
484	GCF_001419105	Campylobacter jejuni CVM 41923	CVM 41923	19270	GCA_001419105.1	1.70194	Contig	50	2	1	12	3	2	1	5
485	GCF_001419125	Campylobacter jejuni CVM 41927	CVM 41927	19270	GCA_001419125.1	1.67247	Contig	8180	181	53	27	10	11	3	6
486	GCF_001419165	Campylobacter jejuni CVM 41933	CVM 41933	19270	GCA_001419165.1	1.72797	Contig	2109	4	7	10	4	10	7	1
487	GCF_001419185	Campylobacter jejuni CVM 41934	CVM 41934	19270	GCA_001419185.1	1.68139	Contig	2109	4	7	10	4	10	7	1
488	GCF_001419195	Campylobacter jejuni CVM 41936	CVM 41936	19270	GCA_001419195.1	1.70274	Contig	2109	4	7	10	4	10	7	1
489	GCF_001419245	Campylobacter jejuni CVM 41943	CVM 41943	19270	GCA_001419245.1	1.65835	Contig	6091	2	4	5	25	11	203	5
490	GCF_001419295	Campylobacter jejuni CVM 41946	CVM 41946	19270	GCA_001419295.1	1.71345	Contig	5453	1	6	137	176	40	478	3
491	GCF_001419475	Campylobacter jejuni CVM 41973	CVM 41973	19270	GCA_001419475.1	1.71669	Contig	2109	4	7	10	4	10	7	1
492	GCF_001419505	Campylobacter jejuni CVM 41974	CVM 41974	19270	GCA_001419505.1	1.67714	Scaffold	22	1	3	6	4	3	3	3
493	GCF_001419635	Campylobacter jejuni CVM 41985	CVM 41985	19270	GCA_001419635.1	1.69128	Contig	2109	4	7	10	4	10	7	1
494	GCF_001419645	Campylobacter jejuni CVM 41975	CVM 41975	19270	GCA_001419645.1	1.72125	Contig	2109	4	7	10	4	10	7	1
495	GCF_001420435	Campylobacter jejuni	2865	19270	GCA_001420435.1	1.82106	Scaffold	1953	7	17	2	2	86	3	1
496	GCF_001432345	Campylobacter jejuni	T1-21	19270	GCA_001432345.1	1.64871	Chromosome	3579	7	4	27	68	11	1	6
497	GCF_001441735	Campylobacter jejuni	CVM N15870	19270	GCA_001441735.1	1.70059	Contig	8177	3	222	29	250	303	25	35
498	GCF_001441755	Campylobacter jejuni	CVM N1630	19270	GCA_001441755.1	1.69242	Contig	6052	2	4	1	2	22	1	5
499	GCF_001442025	Campylobacter jejuni	CVM N15262	19270	GCA_001442025.1	1.72793	Contig	353	7	17	5	2	10	3	6
500	GCF_001442185	Campylobacter jejuni	CVM N279	19270	GCA_001442185.1	1.78037	Contig	3510	7	17	5	2	13	3	6
501	GCF_001442255	Campylobacter jejuni	CVM N534	19270	GCA_001442255.1	1.72355	Contig	462	7	17	5	2	11	3	6
502	GCF_001442295	Campylobacter jejuni	CVM N9016	19270	GCA_001442295.1	1.68684	Contig	48	2	4	1	2	7	1	5
503	GCF_001457695	Campylobacter jejuni	NCTC11351	19270	GCA_001457695.1	1.76644	Complete Genome	403	10	27	16	19	10	5	7
504	GCF_001506185	Campylobacter jejuni	CJ677CC519	19270	GCA_001506185.1	1.64279	Complete Genome	677	10	81	50	99	120	76	52
505	GCF_001506205	Campylobacter jejuni	CJ677CC002	19270	GCA_001506205.1	1.66977	Complete Genome	677	10	81	50	99	120	76	52
506	GCF_001506225	Campylobacter jejuni	CJ677CC534	19270	GCA_001506225.1	1.63383	Complete Genome	677	10	81	50	99	120	76	52
507	GCF_001506245	Campylobacter jejuni	CJ677CC536	19270	GCA_001506245.1	1.64228	Complete Genome	677	10	81	50	99	120	76	52
508	GCF_001506265	Campylobacter jejuni	CJ677CC073	19270	GCA_001506265.1	1.67255	Complete Genome	794	10	81	50	87	120	76	52
509	GCF_001506285	Campylobacter jejuni	CJ677CC521	19270	GCA_001506285.1	1.66843	Complete Genome	677	10	81	50	99	120	76	52
510	GCF_001506305	Campylobacter jejuni	CJ677CC526	19270	GCA_001506305.1	1.6676	Complete Genome	677	10	81	50	99	120	76	52

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
511	GCF_001506345	Campylobacter jejuni	CJ677CC036	19270	GCA_001506345.1	1.66518	Complete Genome	677	10	81	50	99	120	76	52
512	GCF_001506365	Campylobacter jejuni	CJ677CC524	19270	GCA_001506365.1	1.66833	Complete Genome	677	10	81	50	99	120	76	52
513	GCF_001506385	Campylobacter jejuni	CJ677CC016	19270	GCA_001506385.1	1.63764	Complete Genome	677	10	81	50	99	120	76	52
514	GCF_001506405	Campylobacter jejuni	CJ677CC041	19270	GCA_001506405.1	1.63613	Complete Genome	794	10	81	50	87	120	76	52
515	GCF_001506425	Campylobacter jejuni	CJ677CC535	19270	GCA_001506425.1	1.63369	Complete Genome	677	10	81	50	99	120	76	52
516	GCF_001506445	Campylobacter jejuni	CJ677CC092	19270	GCA_001506445.1	1.63272	Complete Genome	677	10	81	50	99	120	76	52
517	GCF_001506465	Campylobacter jejuni	CJ677CC530	19270	GCA_001506465.1	1.64928	Complete Genome	677	10	81	50	99	120	76	52
518	GCF_001506485	Campylobacter jejuni	CJ677CC532	19270	GCA_001506485.1	1.64304	Complete Genome	677	10	81	50	99	120	76	52
519	GCF_001506505	Campylobacter jejuni	CJ677CC529	19270	GCA_001506505.1	1.63127	Complete Genome	677	10	81	50	99	120	76	52
520	GCF_001506525	Campylobacter jejuni	CJ677CC531	19270	GCA_001506525.1	1.66605	Complete Genome	677	10	81	50	99	120	76	52
521	GCF_001506545	Campylobacter jejuni	CJ677CC062	19270	GCA_001506545.1	1.64498	Complete Genome	677	10	81	50	99	120	76	52
522	GCF_001506565	Campylobacter jejuni	CJ677CC059	19270	GCA_001506565.1	1.63251	Complete Genome	677	10	81	50	99	120	76	52
523	GCF_001506585	Campylobacter jejuni	CJ677CC032	19270	GCA_001506585.1	1.67599	Complete Genome	794	10	81	50	87	120	76	52
524	GCF_001506605	Campylobacter jejuni	CJ677CC033	19270	GCA_001506605.1	1.63411	Complete Genome	794	10	81	50	87	120	76	52
525	GCF_001506625	Campylobacter jejuni	CJ677CC537	19270	GCA_001506625.1	1.63315	Complete Genome	677	10	81	50	99	120	76	52
526	GCF_001506645	Campylobacter jejuni	CJ677CC542	19270	GCA_001506645.1	1.6336	Complete Genome	794	10	81	50	87	120	76	52
527	GCF_001506665	Campylobacter jejuni	CJ677CC528	19270	GCA_001506665.1	1.64519	Complete Genome	677	10	81	50	99	120	76	52
528	GCF_001506685	Campylobacter jejuni	CJ677CC538	19270	GCA_001506685.1	1.67595	Complete Genome	677	10	81	50	99	120	76	52
529	GCF_001506705	Campylobacter jejuni	CJ677CC520	19270	GCA_001506705.1	1.67976	Complete Genome	677	10	81	50	99	120	76	52
530	GCF_001506725	Campylobacter jejuni	CJ677CC014	19270	GCA_001506725.1	1.67123	Complete Genome	677	10	81	50	99	120	76	52
531	GCF_001506745	Campylobacter jejuni	CJ677CC039	19270	GCA_001506745.1	1.63563	Complete Genome	677	10	81	50	99	120	76	52
532	GCF_001506765	Campylobacter jejuni	CJ677CC085	19270	GCA_001506765.1	1.63638	Complete Genome	794	10	81	50	87	120	76	52

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
533	GCF_001506785	Campylobacter jejuni	CJ677CC052	19270	GCA_001506785.1	1.63521	Complete Genome	794	10	81	50	87	120	76	52
534	GCF_001506805	Campylobacter jejuni	CJ677CC527	19270	GCA_001506805.1	1.64394	Complete Genome	677	10	81	50	99	120	76	52
535	GCF_001506825	Campylobacter jejuni	CJ677CC078	19270	GCA_001506825.1	1.63137	Complete Genome	677	10	81	50	99	120	76	52
536	GCF_001506845	Campylobacter jejuni	CJ677CC523	19270	GCA_001506845.1	1.66722	Complete Genome	677	10	81	50	99	120	76	52
537	GCF_001506865	Campylobacter jejuni	CJ677CC540	19270	GCA_001506865.1	1.63649	Complete Genome	794	10	81	50	87	120	76	52
538	GCF_001506885	Campylobacter jejuni	CJ677CC040	19270	GCA_001506885.1	1.64065	Complete Genome	677	10	81	50	99	120	76	52
539	GCF_001506905	Campylobacter jejuni	CJ677CC061	19270	GCA_001506905.1	1.63513	Complete Genome	677	10	81	50	99	120	76	52
540	GCF_001506925	Campylobacter jejuni	CJ677CC539	19270	GCA_001506925.1	1.63622	Complete Genome	794	10	81	50	87	120	76	52
541	GCF_001506945	Campylobacter jejuni	CJ677CC533	19270	GCA_001506945.1	1.65705	Complete Genome	677	10	81	50	99	120	76	52
542	GCF_001506965	Campylobacter jejuni	CJ677CC047	19270	GCA_001506965.1	1.63524	Complete Genome	677	10	81	50	99	120	76	52
543	GCF_001506985	Campylobacter jejuni	CJ677CC058	19270	GCA_001506985.1	1.63261	Complete Genome	677	10	81	50	99	120	76	52
544	GCF_001507005	Campylobacter jejuni	CJ677CC013	19270	GCA_001507005.1	1.64269	Complete Genome	677	10	81	50	99	120	76	52
545	GCF_001507025	Campylobacter jejuni	CJ677CC100	19270	GCA_001507025.1	1.66915	Complete Genome	677	10	81	50	99	120	76	52
546	GCF_001507045	Campylobacter jejuni	CJ677CC522	19270	GCA_001507045.1	1.6425	Complete Genome	677	10	81	50	99	120	76	52
547	GCF_001507065	Campylobacter jejuni	CJ677CC094	19270	GCA_001507065.1	1.64112	Complete Genome	677	10	81	50	99	120	76	52
548	GCF_001507085	Campylobacter jejuni	CJ677CC008	19270	GCA_001507085.1	1.62454	Complete Genome	677	10	81	50	99	120	76	52
549	GCF_001507105	Campylobacter jejuni	CJ677CC541	19270	GCA_001507105.1	1.63419	Complete Genome	794	10	81	50	87	120	76	52
550	GCF_001507125	Campylobacter jejuni	CJ677CC024	19270	GCA_001507125.1	1.66157	Complete Genome	677	10	81	50	99	120	76	52
551	GCF_001507145	Campylobacter jejuni	CJ677CC064	19270	GCA_001507145.1	1.64516	Complete Genome	677	10	81	50	99	120	76	52
552	GCF_001507165	Campylobacter jejuni	CJ677CC525	19270	GCA_001507165.1	1.64204	Complete Genome	677	10	81	50	99	120	76	52
553	GCF_001507185	Campylobacter jejuni	CJ677CC026	19270	GCA_001507185.1	1.63571	Complete Genome	677	10	81	50	99	120	76	52
554	GCF_001507205	Campylobacter jejuni	CJ677CC034	19270	GCA_001507205.1	1.63721	Complete Genome	794	10	81	50	87	120	76	52

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
555	GCF_001507225	Campylobacter jejuni	CJ677CC086	19270	GCA_001507225.1	1.62979	Complete Genome	677	10	81	50	99	120	76	52
556	GCF_001507245	Campylobacter jejuni	CJ677CC095	19270	GCA_001507245.1	1.62891	Complete Genome	677	10	81	50	99	120	76	52
557	GCF_001507265	Campylobacter jejuni	CJ677CC012	19270	GCA_001507265.1	1.63395	Complete Genome	794	10	81	50	87	120	76	52
558	GCF_001516405	Campylobacter jejuni HB-CJGB-QYT	HB-CJGB-QYT	19270	GCA_001516405.1	1.68767	Scaffold	5233	2	17	5	3	2	1	5
559	GCF_001516415	Campylobacter jejuni HB-CJGB-LL	HB-CJGB-LL	19270	GCA_001516415.1	1.68091	Scaffold	51	7	17	2	15	23	3	12
560	GCF_001516425	Campylobacter jejuni HB-CJGB-LXC	HB-CJGB-LXC	19270	GCA_001516425.1	1.61026	Scaffold	22	1	3	6	4	3	3	3
561	GCF_001516435	Campylobacter jejuni HB-CJGB-XWM	HB-CJGB-XWM	19270	GCA_001516435.1	1.67189	Scaffold	2049	8	10	16	2	11	12	6
562	GCF_001516485	Campylobacter jejuni BJ-CJGB96G25	BJ-CJGB96G25	19270	GCA_001516485.1	1.61837	Scaffold	22	1	3	6	4	3	3	3
563	GCF_001516495	Campylobacter jejuni BJ-CJGB95377	BJ-CJGB95377	19270	GCA_001516495.1	1.6137	Scaffold	22	1	3	6	4	3	3	3
564	GCF_001516505	Campylobacter jejuni BJ-CJGB96114	BJ-CJGB96114	19270	GCA_001516505.1	1.61333	Scaffold	22	1	3	6	4	3	3	3
565	GCF_001516515	Campylobacter jejuni BJ-CJGB96299	BJ-CJGB96299	19270	GCA_001516515.1	1.68718	Scaffold	5233	2	17	5	3	2	1	5
566	GCF_001516565	Campylobacter jejuni HB-CJGB-LC	HB-CJGB-LC	19270	GCA_001516565.1	1.65425	Scaffold	45	4	7	10	4	1	7	1
567	GCF_001547595	Campylobacter jejuni	BJ-CJD70	19270	GCA_001547595.1	1.60888	Scaffold	3930	7	78	42	4	11	12	8
568	GCF_001547605	Campylobacter jejuni	BJ-CJD120	19270	GCA_001547605.1	1.73758	Scaffold	1232	7	17	5	10	11	3	6
569	GCF_001547615	Campylobacter jejuni	BJ-CJD63	19270	GCA_001547615.1	1.75041	Scaffold	2328	8	2	2	212	153	253	147
570	GCF_001547625	Campylobacter jejuni	BJ-CJD39	19270	GCA_001547625.1	1.65569	Scaffold	653	9	17	2	2	11	12	6
571	GCF_001547675	Campylobacter jejuni	JL-CJHLIU1-1	19270	GCA_001547675.1	1.80833	Scaffold	2274	9	17	5	10	350	3	3
572	GCF_001563565	Campylobacter jejuni	RM3194	19270	GCA_001563565.1	1.73226	Complete Genome	1471	24	171	2	2	89	59	6
573	GCF_001570705	Campylobacter jejuni	NC05-27	19270	GCA_001570705.1	1.77851	Contig	6	63	34	27	33	45	5	7
574	GCF_001587015	Campylobacter jejuni	OD267	19270	GCA_001587015.1	1.82632	Complete Genome	50	2	1	12	3	2	1	5
575	GCF_001587035	Campylobacter jejuni	WP2202	19270	GCA_001587035.1	1.80145	Complete Genome	50	2	1	12	3	2	1	5
576	GCF_001686905	Campylobacter jejuni subsp. jejuni	RM1285	19270	GCA_001686905.1	1.67511	Complete Genome	22	1	3	6	4	3	3	3
577	GCF_001717625	Campylobacter jejuni subsp. jejuni	14980A	19270	GCA_001717625.1	1.75989	Complete Genome	1839	2	222	29	250	303	25	35
578	GCF_001721965	Campylobacter jejuni subsp. jejuni	MTVDSCj13	19270	GCA_001721965.1	1.68441	Complete Genome	460	24	30	2	2	89	59	6
579	GCF_001721985	Campylobacter jejuni subsp. jejuni	MTVDSCj16	19270	GCA_001721985.1	1.78531	Complete Genome	1911	7	84	5	10	119	178	26

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
580	GCF_900000215	Campylobacter jejuni	Campylobacter jejuni	19270	GCA_900000215.1	1.6526	Scaffold	61	1	4	2	2	6	3	17
581	GCF_900036015	Campylobacter jejuni	4	19270	GCA_900036015.1	1.443	Scaffold	403	10	27	16	19	10	5	7
582	GCA_000144995	Neisseria meningitidis K1207	K1207	19400	GCA_000144995.3	2.14003	Scaffold	11	2	3	4	3	8	4	6
583	GCA_000145015	Neisseria meningitidis S0108	S0108	19400	GCA_000145015.3	2.14581	Scaffold	11	2	3	4	3	8	4	6
584	GCA_000392695	Neisseria meningitidis 2003022	2003022	19400	GCA_000392695.2	2.14465	Contig	7	1	1	2	1	3	2	19
585	GCA_000392715	Neisseria meningitidis NM1482	NM1482	19400	GCA_000392715.2	2.15251	Contig	1287	2	3	4	17	8	4	6
586	GCA_000392735	Neisseria meningitidis NM27	NM27	19400	GCA_000392735.2	2.18716	Contig	1622	10	5	18	175	11	9	17
587	GCA_001703735	Neisseria meningitidis	M22790	19400	GCA_001703735.1	2.16817	Chromosom e	2881	179	7	4	56	26	18	8
588	GCA_900038635	Neisseria meningitidis	2842STDY5881727	19400	GCA_900038635.1	2.15636	Scaffold	42	10	6	9	5	9	6	9
589	GCF_000008805	Neisseria meningitidis MC58	MC58	19400	GCA_000008805.1	2.27236	Complete Genome	74	4	10	5	4	5	3	2
590	GCF_000009105	Neisseria meningitidis Z2491	Z2491	19400	GCA_000009105.1	2.18441	Complete Genome	4	1	3	3	1	4	2	3
591	GCF_000009465	Neisseria meningitidis FAM18	FAM18	19400	GCA_000009465.1	2.19496	Complete Genome	11	2	3	4	3	8	4	6
592	GCF_000014105	Neisseria meningitidis 053442	53442	19400	GCA_000014105.1	2.15342	Complete Genome	4821	222	3	58	275	30	5	255
593	GCF_000026965	Neisseria meningitidis 8013	8013	19400	GCA_000026965.1	2.27755	Complete Genome	177	7	8	10	38	10	1	20
594	GCF_000083565	Neisseria meningitidis alpha14	alpha14	19400	GCA_000083565.1	2.14529	Complete Genome	53	16	2	6	25	17	25	22
595	GCF_000144995	Neisseria meningitidis K1207	K1207	19400	GCA_000144995.3	2.14003	Scaffold	11	2	3	4	3	8	4	6
596	GCF_000145015	Neisseria meningitidis S0108	S0108	19400	GCA_000145015.3	2.14581	Scaffold	11	2	3	4	3	8	4	6
597	GCF_000146655	Neisseria meningitidis ATCC 13091	ATCC 13091	19400	GCA_000146655.1	2.26261	Scaffold	7355	6	5	15	9	221	34	13
598	GCF_000152165	Neisseria meningitidis alpha710	alpha710	19400	GCA_000152165.1	2.24295	Complete Genome	136	27	6	9	3	9	6	16
599	GCF_000185025	Neisseria meningitidis H44/76	H44/76	19400	GCA_000185025.2	2.17038	Contig	32	4	10	5	4	6	3	8
600	GCF_000191185	Neisseria meningitidis N1568	N1568	19400	GCA_000191185.2	2.18023	Contig	751	10	3	15	7	8	41	6
601	GCF_000191205	Neisseria meningitidis OX99.30304	OX99.30304	19400	GCA_000191205.2	2.16261	Contig	8094	9	6	9	9	9	514	9
602	GCF_000191225	Neisseria meningitidis M6190	M6190	19400	GCA_000191225.2	2.21311	Contig	1988	2	3	4	3	11	4	6
603	GCF_000191245	Neisseria meningitidis M13399	M13399	19400	GCA_000191245.2	2.26299	Contig	2976	4	10	15	9	8	11	10
604	GCF_000191265	Neisseria meningitidis M0579	M0579	19400	GCA_000191265.2	2.26004	Contig	43	12	6	9	9	9	6	9
605	GCF_000191285	Neisseria meningitidis ES14902	ES14902	19400	GCA_000191285.2	2.19054	Contig	11	2	3	4	3	8	4	6
606	GCF_000191305	Neisseria meningitidis CU385	CU385	19400	GCA_000191305.2	2.25388	Contig	33	8	10	5	4	6	3	8
607	GCF_000191345	Neisseria meningitidis M01-240013	M01-240013	19400	GCA_000191345.2	2.29839	Contig	1159	4	10	2	5	38	11	1
608	GCF_000191425	Neisseria meningitidis G2136	G2136	19400	GCA_000191425.1	2.18486	Complete Genome	8	2	3	7	2	8	5	2

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
609	GCF_000191445	Neisseria meningitidis H44/76	H44/76	19400	GCA_000191445.1	2.24088	Complete Genome	32	4	10	5	4	6	3	8
610	GCF_000191465	Neisseria meningitidis M01-240149	M01-240149	19400	GCA_000191465.1	2.22352	Complete Genome	41	3	6	9	5	9	6	9
611	GCF_000191485	Neisseria meningitidis M01-240355	M01-240355	19400	GCA_000191485.1	2.28778	Complete Genome	213	7	5	1	13	36	53	15
612	GCF_000191505	Neisseria meningitidis M04-240196	M04-240196	19400	GCA_000191505.1	2.25045	Complete Genome	269	4	10	15	9	8	11	9
613	GCF_000191525	Neisseria meningitidis NZ-05/33	NZ-05/33	19400	GCA_000191525.1	2.24897	Complete Genome	42	10	6	9	5	9	6	9
614	GCF_000193815	Neisseria meningitidis NS44	NS44	19400	GCA_000193815.2	2.02209	Contig	23	10	5	18	9	11	9	17
615	GCF_000240565	Neisseria meningitidis Nm3127	Nm3127	19400	GCA_000240565.2	2.07878	Contig	3980	2	7	6	8	9	7	8
616	GCF_000240585	Neisseria meningitidis Nm2732	Nm2732	19400	GCA_000240585.2	2.16496	Contig	22	11	5	18	8	11	24	21
617	GCF_000240605	Neisseria meningitidis Nm6938	Nm6938	19400	GCA_000240605.2	2.17484	Contig	22	11	5	18	8	11	24	21
618	GCF_000240625	Neisseria meningitidis Nm6756	Nm6756	19400	GCA_000240625.2	2.0574	Contig	23	10	5	18	9	11	9	17
619	GCF_000240645	Neisseria meningitidis Nm8663	Nm8663	19400	GCA_000240645.2	2.05789	Contig	23	10	5	18	9	11	9	17
620	GCF_000240665	Neisseria meningitidis Nm1140	Nm1140	19400	GCA_000240665.2	2.05608	Contig	1136	5	4	38	15	22	40	13
621	GCF_000242735	Neisseria meningitidis NM233	NM233	19400	GCA_000242735.2	2.00471	Contig	1621	10	5	182	9	11	9	17
622	GCF_000242755	Neisseria meningitidis NM220	NM220	19400	GCA_000242755.2	2.01769	Contig	23	10	5	18	9	11	9	17
623	GCF_000253215	Neisseria meningitidis WUE 2594	WUE 2594	19400	GCA_000253215.1	2.22725	Complete Genome	5	1	1	2	1	3	2	3
624	GCF_000265135	Neisseria meningitidis DE9686	DE9686	19400	GCA_000265135.1	2.1419	Contig	42	10	6	9	5	9	6	9
625	GCF_000265155	Neisseria meningitidis DE9622	DE9622	19400	GCA_000265155.1	2.13762	Contig	42	10	6	9	5	9	6	9
626	GCF_000265175	Neisseria meningitidis DE9938	DE9938	19400	GCA_000265175.1	2.13553	Contig	42	10	6	9	5	9	6	9
627	GCF_000293245	Neisseria meningitidis 93004	93004	19400	GCA_000293245.1	2.33639	Contig	5594	21	18	6	37	26	15	20
628	GCF_000293265	Neisseria meningitidis 93003	93003	19400	GCA_000293265.1	2.23895	Contig	4959	300	231	53	24	30	8	3
629	GCF_000293285	Neisseria meningitidis NM255	NM255	19400	GCA_000293285.1	2.19637	Contig	2980	6	7	4	5	246	18	8
630	GCF_000293305	Neisseria meningitidis NM140	NM140	19400	GCA_000293305.1	2.16477	Contig	2981	8	4	6	9	5	18	241
631	GCF_000293325	Neisseria meningitidis NM183	NM183	19400	GCA_000293325.1	2.18909	Contig	5467	8	4	6	9	3	18	241
632	GCF_000293345	Neisseria meningitidis NM2781	NM2781	19400	GCA_000293345.1	2.16972	Contig	6937	8	4	6	17	5	18	241
633	GCF_000293365	Neisseria meningitidis 69166	69166	19400	GCA_000293365.1	2.2187	Contig	1	1	3	1	1	1	1	3
634	GCF_000293385	Neisseria meningitidis NM576	NM576	19400	GCA_000293385.1	2.18975	Contig	5467	8	4	6	9	3	18	241
635	GCF_000293405	Neisseria meningitidis 98008	98008	19400	GCA_000293405.2	2.30388	Contig	461	12	5	12	35	60	22	17
636	GCF_000293425	Neisseria meningitidis 80179	80179	19400	GCA_000293425.1	2.18772	Contig	178	7	16	55	10	3	56	46
637	GCF_000293445	Neisseria meningitidis 92045	92045	19400	GCA_000293445.1	2.22392	Contig	4122	12	16	34	17	9	38	17
638	GCF_000293465	Neisseria meningitidis NM2657	NM2657	19400	GCA_000293465.1	2.15449	Contig	60	17	5	19	17	3	26	2
639	GCF_000293625	Neisseria meningitidis NM2795	NM2795	19400	GCA_000293625.1	2.15499	Contig	198	5	4	17	15	14	7	12
640	GCF_000293645	Neisseria meningitidis NM3081	NM3081	19400	GCA_000293645.1	2.34663	Contig	6173	47	3	6	152	143	24	95

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
641	GCF_000293665	Neisseria meningitidis NM3001	NM3001	19400	GCA_000293665.1	2.15248	Contig	1157	8	25	7	17	21	26	49
642	GCF_000304435	Neisseria meningitidis alpha704	alpha704	19400	GCA_000304435.1	1.98336	Contig	198	5	4	17	15	14	7	12
643	GCF_000327545	Neisseria meningitidis 63006	63006	19400	GCA_000327545.2	2.19794	Contig	4	1	3	3	1	4	2	3
644	GCF_000327565	Neisseria meningitidis NM3642	NM3642	19400	GCA_000327565.2	2.21135	Contig	4789	1	1	2	1	3	334	19
645	GCF_000327585	Neisseria meningitidis 97020	97020	19400	GCA_000327585.2	2.19906	Contig	580	2	1	2	1	3	2	3
646	GCF_000327605	Neisseria meningitidis 96023	96023	19400	GCA_000327605.2	2.2096	Contig	5	1	1	2	1	3	2	3
647	GCF_000327625	Neisseria meningitidis M7124	M7124	19400	GCA_000327625.2	2.19069	Contig	11	2	3	4	3	8	4	6
648	GCF_000327665	Neisseria meningitidis 63041	63041	19400	GCA_000327665.2	2.1841	Contig	4	1	3	3	1	4	2	3
649	GCF_000327685	Neisseria meningitidis 97021	97021	19400	GCA_000327685.2	2.19134	Contig	181	10	3	15	7	5	41	31
650	GCF_000327705	Neisseria meningitidis 87255	87255	19400	GCA_000327705.2	2.15821	Contig	10308	2	3	7	10	8	15	6
651	GCF_000327725	Neisseria meningitidis 2007056	2007056	19400	GCA_000327725.2	2.19868	Contig	2859	1	3	2	1	3	2	19
652	GCF_000327745	Neisseria meningitidis 69096	69096	19400	GCA_000327745.2	2.24935	Contig	1	1	3	1	1	1	1	3
653	GCF_000327765	Neisseria meningitidis 61103	61103	19400	GCA_000327765.2	2.19975	Contig	9580	1	3	1	1	1	630	3
654	GCF_000327785	Neisseria meningitidis 65014	65014	19400	GCA_000327785.2	2.17937	Contig	4	1	3	3	1	4	2	3
655	GCF_000327805	Neisseria meningitidis 63049	63049	19400	GCA_000327805.2	2.19532	Contig	4	1	3	3	1	4	2	3
656	GCF_000327825	Neisseria meningitidis 4119	4119	19400	GCA_000327825.2	2.23614	Contig	639	8	10	5	9	6	3	8
657	GCF_000327845	Neisseria meningitidis 9757	9757	19400	GCA_000327845.2	2.24025	Contig	3597	8	10	5	4	6	3	13
658	GCF_000327865	Neisseria meningitidis 9506	9506	19400	GCA_000327865.2	2.24113	Contig	33	8	10	5	4	6	3	8
659	GCF_000327885	Neisseria meningitidis NM174	NM174	19400	GCA_000327885.2	2.17479	Contig	11	2	3	4	3	8	4	6
660	GCF_000327905	Neisseria meningitidis NM762	NM762	19400	GCA_000327905.2	2.19162	Contig	11	2	3	4	3	8	4	6
661	GCF_000327925	Neisseria meningitidis NM586	NM586	19400	GCA_000327925.2	2.22725	Contig	11	2	3	4	3	8	4	6
662	GCF_000327945	Neisseria meningitidis M13255	M13255	19400	GCA_000327945.2	2.25467	Contig	32	4	10	5	4	6	3	8
663	GCF_000327965	Neisseria meningitidis 2006087	2006087	19400	GCA_000327965.2	2.20059	Contig	5789	10	4	15	7	5	41	31
664	GCF_000327985	Neisseria meningitidis 88050	88050	19400	GCA_000327985.2	2.2044	Contig	5	1	1	2	1	3	2	3
665	GCF_000328005	Neisseria meningitidis 98080	98080	19400	GCA_000328005.2	2.18896	Contig	658	2	3	4	3	8	110	20
666	GCF_000328025	Neisseria meningitidis 70030	70030	19400	GCA_000328025.2	2.25676	Contig	1	1	3	1	1	1	1	3
667	GCF_000328045	Neisseria meningitidis 77221	77221	19400	GCA_000328045.2	2.27205	Contig	8798	47	3	58	433	21	59	8
668	GCF_000328065	Neisseria meningitidis 2001212	2001212	19400	GCA_000328065.2	2.21698	Contig	7	1	1	2	1	3	2	19
669	GCF_000328085	Neisseria meningitidis NM3652	NM3652	19400	GCA_000328085.2	2.19374	Contig	8428	510	1	2	1	3	334	19
670	GCF_000328105	Neisseria meningitidis 2004090	2004090	19400	GCA_000328105.2	2.20871	Contig	7	1	1	2	1	3	2	19
671	GCF_000328125	Neisseria meningitidis 12888	12888	19400	GCA_000328125.2	2.26571	Contig	639	8	10	5	9	6	3	8
672	GCF_000328145	Neisseria meningitidis NM126	NM126	19400	GCA_000328145.2	2.21312	Contig	11	2	3	4	3	8	4	6
673	GCF_000328165	Neisseria meningitidis M7089	M7089	19400	GCA_000328165.2	2.16865	Contig	11	2	3	4	3	8	4	6
674	GCF_000328185	Neisseria meningitidis NM418	NM418	19400	GCA_000328185.2	2.29861	Contig	32	4	10	5	4	6	3	8
675	GCF_000328205	Neisseria meningitidis 2002038	2002038	19400	GCA_000328205.2	2.19621	Contig	181	10	3	15	7	5	41	31

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
676	GCF_000328225	Neisseria meningitidis 70012	70012	19400	GCA_000328225.2	2.24121	Contig	1	1	3	1	1	1	1	3
677	GCF_000328245	Neisseria meningitidis 68094	68094	19400	GCA_000328245.2	2.20681	Contig	1	1	3	1	1	1	1	3
678	GCF_000328265	Neisseria meningitidis NM422	NM422	19400	GCA_000328265.2	2.29354	Contig	32	4	10	5	4	6	3	8
679	GCF_000334175	Neisseria meningitidis Nm11003	Nm11003	19400	GCA_000334175.1	2.12743	Scaffold	4821	222	3	58	275	30	5	255
680	GCF_000367485	Neisseria meningitidis NMB	NMB	19400	GCA_000367485.1	2.10294	Contig	4609	2	3	175	2	8	160	2
681	GCF_000386025	Neisseria meningitidis NM115	NM115	19400	GCA_000386025.2	2.16704	Contig	23	10	5	18	9	11	9	17
682	GCF_000386045	Neisseria meningitidis NM90	NM90	19400	GCA_000386045.2	2.25503	Contig	1622	10	5	18	175	11	9	17
683	GCF_000386065	Neisseria meningitidis NM3042	NM3042	19400	GCA_000386065.2	2.17767	Contig	23	10	5	18	9	11	9	17
684	GCF_000386085	Neisseria meningitidis 69155	69155	19400	GCA_000386085.2	2.21037	Contig	8973	1	3	1	25	1	1	3
685	GCF_000386105	Neisseria meningitidis 69176	69176	19400	GCA_000386105.2	2.20253	Contig	8973	1	3	1	25	1	1	3
686	GCF_000386125	Neisseria meningitidis 70021	70021	19400	GCA_000386125.2	2.21186	Contig	1	1	3	1	1	1	1	3
687	GCF_000386145	Neisseria meningitidis 2000080	2000080	19400	GCA_000386145.2	2.15835	Contig	7	1	1	2	1	3	2	19
688	GCF_000386165	Neisseria meningitidis 96060	96060	19400	GCA_000386165.2	2.24024	Contig	1	1	3	1	1	1	1	3
689	GCF_000386205	Neisseria meningitidis 75643	75643	19400	GCA_000386205.2	2.13593	Contig	5	1	1	2	1	3	2	3
690	GCF_000386225	Neisseria meningitidis 75689	75689	19400	GCA_000386225.2	2.19477	Contig	5	1	1	2	1	3	2	3
691	GCF_000386245	Neisseria meningitidis 69100	69100	19400	GCA_000386245.2	2.21163	Contig	1	1	3	1	1	1	1	3
692	GCF_000386265	Neisseria meningitidis 63023	63023	19400	GCA_000386265.2	2.24858	Contig	1	1	3	1	1	1	1	3
693	GCF_000386285	Neisseria meningitidis 61106	61106	19400	GCA_000386285.2	2.24327	Scaffold	1	1	3	1	1	1	1	3
694	GCF_000386305	Neisseria meningitidis 70082	70082	19400	GCA_000386305.2	2.21215	Contig	1	1	3	1	1	1	1	3
695	GCF_000386325	Neisseria meningitidis 65012	65012	19400	GCA_000386325.2	2.17518	Contig	4	1	3	3	1	4	2	3
696	GCF_000386345	Neisseria meningitidis 64182	64182	19400	GCA_000386345.2	2.19052	Contig	4	1	3	3	1	4	2	3
697	GCF_000386365	Neisseria meningitidis 97027	97027	19400	GCA_000386365.2	2.17767	Contig	4	1	3	3	1	4	2	3
698	GCF_000386385	Neisseria meningitidis 96024	96024	19400	GCA_000386385.2	2.14265	Contig	5	1	1	2	1	3	2	3
699	GCF_000386405	Neisseria meningitidis 97008	97008	19400	GCA_000386405.2	2.19749	Contig	5	1	1	2	1	3	2	3
700	GCF_000386425	Neisseria meningitidis 98005	98005	19400	GCA_000386425.2	2.1473	Contig	5	1	1	2	1	3	2	3
701	GCF_000386445	Neisseria meningitidis 2004085	2004085	19400	GCA_000386445.2	2.18555	Contig	7	1	1	2	1	3	2	19
702	GCF_000386465	Neisseria meningitidis 2000063	2000063	19400	GCA_000386465.2	2.19809	Contig	7	1	1	2	1	3	2	19
703	GCF_000386485	Neisseria meningitidis 2002007	2002007	19400	GCA_000386485.2	2.19411	Contig	7	1	1	2	1	3	2	19
704	GCF_000386585	Neisseria meningitidis NM3222	NM3222	19400	GCA_000386585.2	2.18567	Contig	23	10	5	18	9	11	9	17
705	GCF_000386605	Neisseria meningitidis NM3131	NM3131	19400	GCA_000386605.2	2.13632	Contig	893	10	5	18	9	26	9	17
706	GCF_000386625	Neisseria meningitidis NM3144	NM3144	19400	GCA_000386625.2	2.1515	Contig	6799	10	288	18	9	11	9	17
707	GCF_000386645	Neisseria meningitidis NM3158	NM3158	19400	GCA_000386645.2	2.12791	Contig	6800	10	5	18	9	11	9	2
708	GCF_000386665	Neisseria meningitidis NM3164	NM3164	19400	GCA_000386665.2	2.23558	Contig	3582	10	5	314	9	11	9	17
709	GCF_000386685	Neisseria meningitidis NM51	NM51	19400	GCA_000386685.2	2.11328	Contig	23	10	5	18	9	11	9	17
710	GCF_000386705	Neisseria meningitidis NM43	NM43	19400	GCA_000386705.2	2.12367	Contig	11	2	3	4	3	8	4	6

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
711	GCF_000386725	Neisseria meningitidis NM1495	NM1495	19400	GCA_000386725.2	2.13443	Contig	11	2	3	4	3	8	4	6
712	GCF_000386745	Neisseria meningitidis 73696	73696	19400	GCA_000386745.2	2.26003	Contig	8813	8	5	2	5	26	17	588
713	GCF_000386765	Neisseria meningitidis 73704	73704	19400	GCA_000386765.2	2.10715	Contig	11	2	3	4	3	8	4	6
714	GCF_000386785	Neisseria meningitidis 81858	81858	19400	GCA_000386785.2	2.28923	Contig	344	7	5	2	72	3	75	21
715	GCF_000386805	Neisseria meningitidis 2002020	2002020	19400	GCA_000386805.2	2.20174	Contig	32	4	10	5	4	6	3	8
716	GCF_000386825	Neisseria meningitidis NM477	NM477	19400	GCA_000386825.2	2.28741	Contig	32	4	10	5	4	6	3	8
717	GCF_000386845	Neisseria meningitidis M13265	M13265	19400	GCA_000386845.2	2.156	Contig	32	4	10	5	4	6	3	8
718	GCF_000386865	Neisseria meningitidis NM3147	NM3147	19400	GCA_000386865.2	2.25145	Contig	11	2	3	4	3	8	4	6
719	GCF_000386885	Neisseria meningitidis 2001072	2001072	19400	GCA_000386885.2	2.15231	Contig	11	2	3	4	3	8	4	6
720	GCF_000386945	Neisseria meningitidis 2001001	2001001	19400	GCA_000386945.2	2.19368	Contig	167	2	7	6	17	16	18	8
721	GCF_000386965	Neisseria meningitidis 2004032	2004032	19400	GCA_000386965.2	2.13412	Contig	2881	179	7	4	56	26	18	8
722	GCF_000386985	Neisseria meningitidis 2001213	2001213	19400	GCA_000386985.2	2.13988	Contig	11	2	3	4	3	8	4	6
723	GCF_000387005	Neisseria meningitidis 2004264	2004264	19400	GCA_000387005.2	2.21834	Contig	11	2	3	4	3	8	4	6
724	GCF_000387025	Neisseria meningitidis 2000175	2000175	19400	GCA_000387025.2	2.22809	Contig	11	2	3	4	3	8	4	6
725	GCF_000387045	Neisseria meningitidis 2005079	2005079	19400	GCA_000387045.2	2.11614	Contig	11	2	3	4	3	8	4	6
726	GCF_000387065	Neisseria meningitidis 2005040	2005040	19400	GCA_000387065.2	2.15067	Contig	11	2	3	4	3	8	4	6
727	GCF_000387085	Neisseria meningitidis 2002004	2002004	19400	GCA_000387085.2	2.25148	Contig	11	2	3	4	3	8	4	6
728	GCF_000387105	Neisseria meningitidis 2005172	2005172	19400	GCA_000387105.2	2.17793	Contig	181	10	3	15	7	5	41	31
729	GCF_000387125	Neisseria meningitidis 2008223	2008223	19400	GCA_000387125.2	2.17506	Contig	181	10	3	15	7	5	41	31
730	GCF_000387165	Neisseria meningitidis NM271	NM271	19400	GCA_000387165.2	2.18201	Scaffold	23	10	5	18	9	11	9	17
731	GCF_000387225	Neisseria meningitidis NM313	NM313	19400	GCA_000387225.2	2.17961	Contig	11	2	3	4	3	8	4	6
732	GCF_000387265	Neisseria meningitidis NM80	NM80	19400	GCA_000387265.2	2.11674	Contig	1621	10	5	182	9	11	9	17
733	GCF_000387285	Neisseria meningitidis NM165	NM165	19400	GCA_000387285.2	2.11121	Contig	23	10	5	18	9	11	9	17
734	GCF_000387305	Neisseria meningitidis NM3223	NM3223	19400	GCA_000387305.2	2.12085	Contig	23	10	5	18	9	11	9	17
735	GCF_000387345	Neisseria meningitidis NM32	NM32	19400	GCA_000387345.2	2.22152	Contig	11	2	3	4	3	8	4	6
736	GCF_000387365	Neisseria meningitidis NM35	NM35	19400	GCA_000387365.2	2.17073	Contig	11	2	3	4	3	8	4	6
737	GCF_000387385	Neisseria meningitidis NM36	NM36	19400	GCA_000387385.2	2.10289	Contig	11	2	3	4	3	8	4	6
738	GCF_000392355	Neisseria meningitidis 97018	97018	19400	GCA_000392355.1	2.15234	Contig	580	2	1	2	1	3	2	3
739	GCF_000392695	Neisseria meningitidis 2003022	2003022	19400	GCA_000392695.2	2.14465	Contig	7	1	1	2	1	3	2	19
740	GCF_000392715	Neisseria meningitidis NM1482	NM1482	19400	GCA_000392715.2	2.15251	Contig	1287	2	3	4	17	8	4	6
741	GCF_000392735	Neisseria meningitidis NM27	NM27	19400	GCA_000392735.2	2.18716	Contig	1622	10	5	18	175	11	9	17
742	GCF_000413175	Neisseria meningitidis 2001068	2001068	19400	GCA_000413175.2	2.2092	Contig	11	2	3	4	3	8	4	6
743	GCF_000413215	Neisseria meningitidis NM134	NM134	19400	GCA_000413215.2	2.18801	Contig	185	12	5	34	17	5	38	17
744	GCF_000413235	Neisseria meningitidis 98002	98002	19400	GCA_000413235.2	2.19791	Contig	181	10	3	15	7	5	41	31
745	GCF_000448005	Neisseria meningitidis 96037	96037	19400	GCA_000448005.1	2.24187	Contig	291	1	2	9	9	9	6	10

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
746	GCF_000448025	Neisseria meningitidis 2002030	2002030	19400	GCA_000448025.1	2.2098	Contig	32	4	10	5	4	6	3	8
747	GCF_000448065	Neisseria meningitidis NM3139	NM3139	19400	GCA_000448065.1	2.16222	Contig	4682	9	220	9	17	9	6	9
748	GCF_000448085	Neisseria meningitidis NM045	NM045	19400	GCA_000448085.1	2.26455	Contig	136	27	6	9	3	9	6	16
749	GCF_000448125	Neisseria meningitidis NM003	NM003	19400	GCA_000448125.1	2.19688	Contig	437	9	6	9	17	9	6	9
750	GCF_000448145	Neisseria meningitidis NM1476	NM1476	19400	GCA_000448145.1	2.26171	Contig	32	4	10	5	4	6	3	8
751	GCF_000448165	Neisseria meningitidis NM0552	NM0552	19400	GCA_000448165.1	2.26016	Contig	11268	688	6	9	60	9	6	9
752	GCF_000448205	Neisseria meningitidis NM3173	NM3173	19400	GCA_000448205.1	2.20496	Contig	32	4	10	5	4	6	3	8
753	GCF_000448225	Neisseria meningitidis NM3230	NM3230	19400	GCA_000448225.1	2.26808	Contig	154	3	6	9	5	11	6	9
754	GCF_000464995	Neisseria meningitidis LNP27256	LNP27256	19400	GCA_000464995.1	2.17557	Contig	11	2	3	4	3	8	4	6
755	GCF_000471865	Neisseria meningitidis Nm10259	Nm10259	19400	GCA_000471865.1	2.05389	Contig	60	17	5	19	17	3	26	2
756	GCF_000471885	Neisseria meningitidis Nm9418	Nm9418	19400	GCA_000471885.1	2.04673	Contig	60	17	5	19	17	3	26	2
757	GCF_000626595	Neisseria meningitidis	510612	19400	GCA_000626595.1	2.18802	Complete Genome	7	1	1	2	1	3	2	19
758	GCF_000763745	Neisseria meningitidis	2419	19400	GCA_000763745.1	2.19038	Contig	11	2	3	4	3	8	4	6
759	GCF_000785065	Neisseria meningitidis	NM3687	19400	GCA_000785065.1	2.22323	Contig	11	2	3	4	3	8	4	6
760	GCF_000785075	Neisseria meningitidis	NM3681	19400	GCA_000785075.1	2.19896	Contig	11	2	3	4	3	8	4	6
761	GCF_000787195	Neisseria meningitidis	L91543	19400	GCA_000787195.2	2.17319	Complete Genome	11	2	3	4	3	8	4	6
762	GCF_000800235	Neisseria meningitidis	NM3686	19400	GCA_000800235.1	2.19527	Complete Genome	11	2	3	4	3	8	4	6
763	GCF_000800275	Neisseria meningitidis M7124	M7124	19400	GCA_000800275.1	2.17948	Complete Genome	11	2	3	4	3	8	4	6
764	GCF_000800315	Neisseria meningitidis	NM3682	19400	GCA_000800315.1	2.19667	Complete Genome	11	2	3	4	3	8	4	6
765	GCF_000800355	Neisseria meningitidis	NM3683	19400	GCA_000800355.1	2.19922	Complete Genome	11	2	3	4	3	8	4	6
766	GCF_000800415	Neisseria meningitidis	M10208	19400	GCA_000800415.1	2.18323	Complete Genome	11	2	3	4	3	8	4	6
767	GCF_001029815	Neisseria meningitidis	B6116/77	19400	GCA_001029815.1	2.18767	Complete Genome	10	2	3	4	2	8	15	2
768	GCF_001083305	Neisseria meningitidis	NM1845	19400	GCA_001083305.1	2.06758	Scaffold	7	1	1	2	1	3	2	19
769	GCF_001083765	Neisseria meningitidis	NM1573	19400	GCA_001083765.1	2.06317	Scaffold	7	1	1	2	1	3	2	19
770	GCF_001083885	Neisseria meningitidis	NM1549	19400	GCA_001083885.1	2.05669	Scaffold	7	1	1	2	1	3	2	19
771	GCF_001085765	Neisseria meningitidis	NM1672	19400	GCA_001085765.1	2.06946	Scaffold	7	1	1	2	1	3	2	19
772	GCF_001086285	Neisseria meningitidis	NM2700	19400	GCA_001086285.1	2.04436	Scaffold	2859	1	3	2	1	3	2	19
773	GCF_001087685	Neisseria meningitidis	NM2263	19400	GCA_001087685.1	2.03875	Contig	2859	1	3	2	1	3	2	19
774	GCF_001088805	Neisseria meningitidis	NM1361	19400	GCA_001088805.1	2.06799	Scaffold	7	1	1	2	1	3	2	19
775	GCF_001089325	Neisseria meningitidis	NM2254	19400	GCA_001089325.1	2.05249	Scaffold	2859	1	3	2	1	3	2	19

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
776	GCF_001090365	Neisseria meningitidis	NM2206	19400	GCA_001090365.1	2.05274	Scaffold	2859	1	3	2	1	3	2	19
777	GCF_001091005	Neisseria meningitidis	NM3128	19400	GCA_001091005.1	2.06435	Scaffold	2859	1	3	2	1	3	2	19
778	GCF_001092105	Neisseria meningitidis	NM2935	19400	GCA_001092105.1	2.05764	Scaffold	2859	1	3	2	1	3	2	19
779	GCF_001092265	Neisseria meningitidis	NM2813	19400	GCA_001092265.1	2.04054	Contig	2859	1	3	2	1	3	2	19
780	GCF_001093625	Neisseria meningitidis	NM2857	19400	GCA_001093625.1	2.06551	Contig	2859	1	3	2	1	3	2	19
781	GCF_001097185	Neisseria meningitidis	NM2393	19400	GCA_001097185.1	2.05657	Scaffold	2859	1	3	2	1	3	2	19
782	GCF_001098105	Neisseria meningitidis	NM1826	19400	GCA_001098105.1	2.06986	Contig	7	1	1	2	1	3	2	19
783	GCF_001098525	Neisseria meningitidis	NM2811	19400	GCA_001098525.1	2.0391	Scaffold	2859	1	3	2	1	3	2	19
784	GCF_001099965	Neisseria meningitidis	NM2433	19400	GCA_001099965.1	2.05511	Scaffold	2859	1	3	2	1	3	2	19
785	GCF_001100325	Neisseria meningitidis	NM2009	19400	GCA_001100325.1	2.05312	Scaffold	7	1	1	2	1	3	2	19
786	GCF_001100565	Neisseria meningitidis	NM2244	19400	GCA_001100565.1	2.0829	Scaffold	2859	1	3	2	1	3	2	19
787	GCF_001100625	Neisseria meningitidis	NM2237	19400	GCA_001100625.1	2.04629	Contig	2859	1	3	2	1	3	2	19
788	GCF_001101725	Neisseria meningitidis	NM2239	19400	GCA_001101725.1	2.04012	Contig	2859	1	3	2	1	3	2	19
789	GCF_001103205	Neisseria meningitidis	NM2617	19400	GCA_001103205.1	2.04006	Contig	2859	1	3	2	1	3	2	19
790	GCF_001107205	Neisseria meningitidis	NM1758	19400	GCA_001107205.1	2.05136	Contig	7	1	1	2	1	3	2	19
791	GCF_001108785	Neisseria meningitidis	NM1325	19400	GCA_001108785.1	2.06735	Contig	7	1	1	2	1	3	2	19
792	GCF_001112125	Neisseria meningitidis	NM1779	19400	GCA_001112125.1	2.08126	Scaffold	7	1	1	2	1	3	2	19
793	GCF_001112205	Neisseria meningitidis	NM2439	19400	GCA_001112205.1	2.04873	Scaffold	2859	1	3	2	1	3	2	19
794	GCF_001112425	Neisseria meningitidis	NM2602	19400	GCA_001112425.1	2.05183	Scaffold	2859	1	3	2	1	3	2	19
795	GCF_001112805	Neisseria meningitidis	NM2181	19400	GCA_001112805.1	2.04692	Scaffold	2859	1	3	2	1	3	2	19
796	GCF_001113425	Neisseria meningitidis	NM1561	19400	GCA_001113425.1	2.05841	Scaffold	7	1	1	2	1	3	2	19
797	GCF_001114605	Neisseria meningitidis	NM1921	19400	GCA_001114605.1	2.0595	Contig	7	1	1	2	1	3	2	19
798	GCF_001116265	Neisseria meningitidis	NM1757	19400	GCA_001116265.1	2.05494	Scaffold	7	1	1	2	1	3	2	19
799	GCF_001116645	Neisseria meningitidis	NM1928	19400	GCA_001116645.1	2.05826	Scaffold	7	1	1	2	1	3	2	19
800	GCF_001117325	Neisseria meningitidis	NM1919	19400	GCA_001117325.1	2.07095	Scaffold	7	1	1	2	1	3	2	19
801	GCF_001117425	Neisseria meningitidis	NM2381	19400	GCA_001117425.1	2.05554	Scaffold	2859	1	3	2	1	3	2	19
802	GCF_001119825	Neisseria meningitidis	NM2232	19400	GCA_001119825.1	2.03971	Contig	2859	1	3	2	1	3	2	19
803	GCF_001120865	Neisseria meningitidis	NM2335	19400	GCA_001120865.1	2.05325	Scaffold	2859	1	3	2	1	3	2	19
804	GCF_001121905	Neisseria meningitidis	NM1544	19400	GCA_001121905.1	2.05194	Contig	7	1	1	2	1	3	2	19
805	GCF_001122225	Neisseria meningitidis	NM1910	19400	GCA_001122225.1	2.1399	Scaffold	7	1	1	2	1	3	2	19
806	GCF_001123285	Neisseria meningitidis	NM1359	19400	GCA_001123285.1	2.04496	Contig	7	1	1	2	1	3	2	19
807	GCF_001124185	Neisseria meningitidis	NM1666	19400	GCA_001124185.1	2.07799	Scaffold	7	1	1	2	1	3	2	19
808	GCF_001126085	Neisseria meningitidis	NM2369	19400	GCA_001126085.1	2.05732	Contig	2859	1	3	2	1	3	2	19
809	GCF_001126645	Neisseria meningitidis	NM2810	19400	GCA_001126645.1	2.03931	Contig	2859	1	3	2	1	3	2	19
810	GCF_001127705	Neisseria meningitidis	NM1363	19400	GCA_001127705.1	2.06581	Scaffold	7	1	1	2	1	3	2	19

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
811	GCF_001127825	Neisseria meningitidis	NM2228	19400	GCA_001127825.1	2.0468	Scaffold	2859	1	3	2	1	3	2	19
812	GCF_001128105	Neisseria meningitidis	NM1892	19400	GCA_001128105.1	2.05614	Scaffold	7	1	1	2	1	3	2	19
813	GCF_001128885	Neisseria meningitidis	NM2025	19400	GCA_001128885.1	2.0509	Contig	7	1	1	2	1	3	2	19
814	GCF_001131325	Neisseria meningitidis	NM2187	19400	GCA_001131325.1	2.06029	Scaffold	2859	1	3	2	1	3	2	19
815	GCF_001134645	Neisseria meningitidis	NM1937	19400	GCA_001134645.1	2.06315	Scaffold	7	1	1	2	1	3	2	19
816	GCF_001134725	Neisseria meningitidis	NM2717	19400	GCA_001134725.1	2.05716	Scaffold	2859	1	3	2	1	3	2	19
817	GCF_001136445	Neisseria meningitidis	NM1891	19400	GCA_001136445.1	2.06511	Scaffold	7	1	1	2	1	3	2	19
818	GCF_001137985	Neisseria meningitidis	NM1893	19400	GCA_001137985.1	2.05319	Scaffold	7	1	1	2	1	3	2	19
819	GCF_001139145	Neisseria meningitidis	NM2389	19400	GCA_001139145.1	2.04271	Scaffold	2859	1	3	2	1	3	2	19
820	GCF_001139505	Neisseria meningitidis	NM1976	19400	GCA_001139505.1	2.05405	Scaffold	7	1	1	2	1	3	2	19
821	GCF_001143785	Neisseria meningitidis	NM2193	19400	GCA_001143785.1	2.05685	Contig	2859	1	3	2	1	3	2	19
822	GCF_001145125	Neisseria meningitidis	NM2382	19400	GCA_001145125.1	2.05212	Scaffold	2859	1	3	2	1	3	2	19
823	GCF_001147245	Neisseria meningitidis	NM2008	19400	GCA_001147245.1	2.05185	Scaffold	7	1	1	2	1	3	2	19
824	GCF_001148025	Neisseria meningitidis	NM2264	19400	GCA_001148025.1	2.05606	Scaffold	2859	1	3	2	1	3	2	19
825	GCF_001148685	Neisseria meningitidis	NM1797	19400	GCA_001148685.1	2.07488	Contig	7	1	1	2	1	3	2	19
826	GCF_001151925	Neisseria meningitidis	NM2524	19400	GCA_001151925.1	2.0723	Scaffold	2859	1	3	2	1	3	2	19
827	GCF_001152085	Neisseria meningitidis	NM1931	19400	GCA_001152085.1	2.05697	Contig	7	1	1	2	1	3	2	19
828	GCF_001152665	Neisseria meningitidis	NM1837	19400	GCA_001152665.1	2.05545	Contig	7	1	1	2	1	3	2	19
829	GCF_001154845	Neisseria meningitidis	NM2606	19400	GCA_001154845.1	2.03273	Contig	2859	1	3	2	1	3	2	19
830	GCF_001155085	Neisseria meningitidis	NM1895	19400	GCA_001155085.1	2.0531	Scaffold	7	1	1	2	1	3	2	19
831	GCF_001158305	Neisseria meningitidis	NM2441	19400	GCA_001158305.1	2.03112	Scaffold	2859	1	3	2	1	3	2	19
832	GCF_001159185	Neisseria meningitidis	NM1446	19400	GCA_001159185.1	2.04811	Scaffold	7	1	1	2	1	3	2	19
833	GCF_001160445	Neisseria meningitidis	NM1578	19400	GCA_001160445.1	2.05335	Scaffold	7	1	1	2	1	3	2	19
834	GCF_001161625	Neisseria meningitidis	NM1360	19400	GCA_001161625.1	2.05437	Contig	7	1	1	2	1	3	2	19
835	GCF_001162925	Neisseria meningitidis	NM1550	19400	GCA_001162925.1	2.06402	Contig	7	1	1	2	1	3	2	19
836	GCF_001163225	Neisseria meningitidis	NM1362	19400	GCA_001163225.1	2.06508	Contig	7	1	1	2	1	3	2	19
837	GCF_001163625	Neisseria meningitidis	NM1963	19400	GCA_001163625.1	2.05151	Contig	7	1	1	2	1	3	2	19
838	GCF_001165245	Neisseria meningitidis	NM2033	19400	GCA_001165245.1	2.24626	Scaffold	7	1	1	2	1	3	2	19
839	GCF_001165865	Neisseria meningitidis	NM2032	19400	GCA_001165865.1	2.05028	Contig	7	1	1	2	1	3	2	19
840	GCF_001166965	Neisseria meningitidis	NM1831	19400	GCA_001166965.1	2.05363	Contig	7	1	1	2	1	3	2	19
841	GCF_001169725	Neisseria meningitidis	NM2332	19400	GCA_001169725.1	2.04253	Scaffold	2859	1	3	2	1	3	2	19
842	GCF_001183075	Neisseria meningitidis	341215	19400	GCA_001183075.1	2.1219	Scaffold	4821	222	3	58	275	30	5	255
843	GCF_001183095	Neisseria meningitidis	220601	19400	GCA_001183095.1	2.11579	Scaffold	4821	222	3	58	275	30	5	255
844	GCF_001183125	Neisseria meningitidis	440902	19400	GCA_001183125.1	2.11964	Scaffold	4821	222	3	58	275	30	5	255
845	GCF_001183135	Neisseria meningitidis	100603	19400	GCA_001183135.1	2.14193	Scaffold	4821	222	3	58	275	30	5	255

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
846	GCF_001183165	Neisseria meningitidis	131148	19400	GCA_001183165.1	2.08342	Scaffold	9936	222	3	58	275	30	5	12
847	GCF_001183175	Neisseria meningitidis	130803	19400	GCA_001183175.1	2.11992	Scaffold	6928	222	3	58	7	30	5	255
848	GCF_001183205	Neisseria meningitidis	340552	19400	GCA_001183205.1	2.11326	Scaffold	4897	222	3	58	275	30	18	8
849	GCF_001183215	Neisseria meningitidis	34173	19400	GCA_001183215.1	2.06854	Scaffold	9477	222	3	58	275	386	18	13
850	GCF_001183245	Neisseria meningitidis	100572	19400	GCA_001183245.1	2.1494	Scaffold	5610	222	3	58	372	30	18	17
851	GCF_001183255	Neisseria meningitidis	440501	19400	GCA_001183255.1	2.09925	Scaffold	4831	8	3	4	275	30	5	255
852	GCF_001183285	Neisseria meningitidis	100514	19400	GCA_001183285.1	2.03491	Scaffold	4832	222	3	58	275	6	11	255
853	GCF_001183295	Neisseria meningitidis	421007	19400	GCA_001183295.1	2.14778	Scaffold	4821	222	3	58	275	30	5	255
854	GCF_001183325	Neisseria meningitidis	420718	19400	GCA_001183325.1	2.0924	Scaffold	11920	751	3	58	275	30	5	255
855	GCF_001183335	Neisseria meningitidis	100601	19400	GCA_001183335.1	2.15403	Scaffold	10737	610	3	58	92	30	5	255
856	GCF_001183365	Neisseria meningitidis	360624	19400	GCA_001183365.1	2.12149	Scaffold	5473	222	3	374	4	26	5	255
857	GCF_001183375	Neisseria meningitidis	311112	19400	GCA_001183375.1	2.09819	Scaffold	5798	222	3	58	275	263	5	255
858	GCF_001183405	Neisseria meningitidis	100530	19400	GCA_001183405.1	2.14805	Scaffold	5614	222	3	58	261	263	5	2
859	GCF_001183415	Neisseria meningitidis	321114	19400	GCA_001183415.1	2.15649	Scaffold	3200	222	3	58	261	263	5	255
860	GCF_001183445	Neisseria meningitidis	370601	19400	GCA_001183445.1	2.1522	Scaffold	3200	222	3	58	261	263	5	255
861	GCF_001183465	Neisseria meningitidis	330505	19400	GCA_001183465.1	2.1268	Contig	4821	222	3	58	275	30	5	255
862	GCF_001183485	Neisseria meningitidis	320501	19400	GCA_001183485.1	2.12723	Scaffold	4820	3	3	58	275	30	5	255
863	GCF_001183495	Neisseria meningitidis	320503	19400	GCA_001183495.1	2.12185	Scaffold	4821	222	3	58	275	30	5	255
864	GCF_001349075	Neisseria meningitidis	NM2431	19400	GCA_001349075.1	2.04014	Scaffold	2859	1	3	2	1	3	2	19
865	GCF_001349095	Neisseria meningitidis	NM3129	19400	GCA_001349095.1	2.06368	Scaffold	2859	1	3	2	1	3	2	19
866	GCF_001349115	Neisseria meningitidis	NM2933	19400	GCA_001349115.1	2.10509	Scaffold	2859	1	3	2	1	3	2	19
867	GCF_001349135	Neisseria meningitidis	NM2188	19400	GCA_001349135.1	2.04331	Contig	2859	1	3	2	1	3	2	19
868	GCF_001349155	Neisseria meningitidis	NM2856	19400	GCA_001349155.1	2.05729	Scaffold	2859	1	3	2	1	3	2	19
869	GCF_001349175	Neisseria meningitidis	NM1805	19400	GCA_001349175.1	2.06259	Scaffold	7	1	1	2	1	3	2	19
870	GCF_001349195	Neisseria meningitidis	NM1364	19400	GCA_001349195.1	2.12056	Scaffold	7	1	1	2	1	3	2	19
871	GCF_001349215	Neisseria meningitidis	NM1938	19400	GCA_001349215.1	2.05604	Contig	7	1	1	2	1	3	2	19
872	GCF_001349235	Neisseria meningitidis	NM2718	19400	GCA_001349235.1	2.08323	Scaffold	2859	1	3	2	1	3	2	19
873	GCF_001349295	Neisseria meningitidis	NM1673	19400	GCA_001349295.1	2.06146	Scaffold	7	1	1	2	1	3	2	19
874	GCF_001349315	Neisseria meningitidis	NM2934	19400	GCA_001349315.1	2.06244	Scaffold	2859	1	3	2	1	3	2	19
875	GCF_001349335	Neisseria meningitidis	NM1264	19400	GCA_001349335.1	2.07552	Scaffold	7	1	1	2	1	3	2	19
876	GCF_001349355	Neisseria meningitidis	NM1482	19400	GCA_001349355.1	2.05757	Scaffold	7	1	1	2	1	3	2	19
877	GCF_001407225	Neisseria meningitidis	M20599	19400	GCA_001407225.1	2.25841	Contig	11	2	3	4	3	8	4	6
878	GCF_001407245	Neisseria meningitidis	M1412	19400	GCA_001407245.1	2.11037	Contig	11	2	3	4	3	8	4	6
879	GCF_001630495	Neisseria meningitidis	87Mo	19400	GCA_001630495.1	2.11359	Contig	11	2	3	4	3	8	4	6
880	GCF_001630505	Neisseria meningitidis	85Mo	19400	GCA_001630505.1	2.1	Contig	11	2	3	4	3	8	4	6

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
881	GCF_001697105	Neisseria meningitidis	DE8669	19400	GCA_001697105.1	2.2301	Complete Genome	42	10	6	9	5	9	6	9
882	GCF_001697125	Neisseria meningitidis	DE10444	19400	GCA_001697125.1	2.17062	Complete Genome	23	10	5	18	9	11	9	17
883	GCF_001697165	Neisseria meningitidis	DE8555	19400	GCA_001697165.1	2.20793	Complete Genome	11	2	3	4	3	8	4	6
884	GCF_001697205	Neisseria meningitidis	WUE2121	19400	GCA_001697205.1	2.20685	Complete Genome	11	2	3	4	3	8	4	6
885	GCF_001697325	Neisseria meningitidis	M12752	19400	GCA_001697325.1	2.17388	Complete Genome	11	2	3	4	3	8	4	6
886	GCF_001697345	Neisseria meningitidis	M22819	19400	GCA_001697345.1	2.17369	Complete Genome	2881	179	7	4	56	26	18	8
887	GCF_001697365	Neisseria meningitidis	M22809	19400	GCA_001697365.1	2.18217	Complete Genome	2881	179	7	4	56	26	18	8
888	GCF_001697385	Neisseria meningitidis	M09293	19400	GCA_001697385.1	2.16151	Complete Genome	11	2	3	4	3	8	4	6
889	GCF_001697405	Neisseria meningitidis	M22189	19400	GCA_001697405.1	2.1727	Complete Genome	11	2	3	4	3	8	4	6
890	GCF_001697425	Neisseria meningitidis	M07149	19400	GCA_001697425.1	2.17351	Complete Genome	11	2	3	4	3	8	4	6
891	GCF_001697445	Neisseria meningitidis	M25474	19400	GCA_001697445.1	2.17258	Complete Genome	11	2	3	4	3	8	4	6
892	GCF_001697465	Neisseria meningitidis	M08001	19400	GCA_001697465.1	2.1622	Complete Genome	11	2	3	4	3	8	4	6
893	GCF_001697485	Neisseria meningitidis	M22748	19400	GCA_001697485.1	2.1575	Complete Genome	11	2	3	4	3	8	4	6
894	GCF_001697505	Neisseria meningitidis	M22811	19400	GCA_001697505.1	2.1857	Complete Genome	2881	179	7	4	56	26	18	8
895	GCF_001697525	Neisseria meningitidis	M22772	19400	GCA_001697525.1	2.17361	Complete Genome	11	2	3	4	3	8	4	6
896	GCF_001697545	Neisseria meningitidis	M22769	19400	GCA_001697545.1	2.1685	Complete Genome	11	2	3	4	3	8	4	6
897	GCF_001697585	Neisseria meningitidis	M24730	19400	GCA_001697585.1	2.17536	Complete Genome	11	2	3	4	3	8	4	6
898	GCF_001697605	Neisseria meningitidis	M22801	19400	GCA_001697605.1	2.17243	Complete Genome	11	2	3	4	3	8	4	6
899	GCF_001697625	Neisseria meningitidis	M22804	19400	GCA_001697625.1	2.17479	Complete Genome	2881	179	7	4	56	26	18	8
900	GCF_001697645	Neisseria meningitidis	M25438	19400	GCA_001697645.1	2.17198	Complete Genome	11	2	3	4	3	8	4	6
901	GCF_001697665	Neisseria meningitidis	M23413	19400	GCA_001697665.1	2.17372	Complete Genome	11	2	3	4	3	8	4	6
902	GCF_001697685	Neisseria meningitidis	M22722	19400	GCA_001697685.1	2.17289	Complete Genome	11	2	3	4	3	8	4	6

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
903	GCF_001697705	Neisseria meningitidis	M25070	19400	GCA_001697705.1	2.168	Complete Genome	11	2	3	4	3	8	4	6
904	GCF_001697725	Neisseria meningitidis	M09261	19400	GCA_001697725.1	2.15434	Complete Genome	11	2	3	4	3	8	4	6
905	GCF_001697745	Neisseria meningitidis	M25462	19400	GCA_001697745.1	2.17404	Complete Genome	11	2	3	4	3	8	4	6
906	GCF_001697765	Neisseria meningitidis	M27559	19400	GCA_001697765.1	2.17374	Complete Genome	11	2	3	4	3	8	4	6
907	GCF_001697785	Neisseria meningitidis	M25472	19400	GCA_001697785.1	2.17015	Complete Genome	11	2	3	4	3	8	4	6
908	GCF_001697805	Neisseria meningitidis	M22759	19400	GCA_001697805.1	2.16831	Complete Genome	11	2	3	4	3	8	4	6
909	GCF_001697825	Neisseria meningitidis	M25087	19400	GCA_001697825.1	2.16799	Complete Genome	11	2	3	4	3	8	4	6
910	GCF_001697845	Neisseria meningitidis	M22783	19400	GCA_001697845.1	2.18057	Complete Genome	2881	179	7	4	56	26	18	8
911	GCF_001697865	Neisseria meningitidis	M22828	19400	GCA_001697865.1	2.17293	Complete Genome	2881	179	7	4	56	26	18	8
912	GCF_001697885	Neisseria meningitidis	M25459	19400	GCA_001697885.1	2.17312	Complete Genome	11	2	3	4	3	8	4	6
913	GCF_001697905	Neisseria meningitidis	M22160	19400	GCA_001697905.1	2.17734	Complete Genome	11	2	3	4	3	8	4	6
914	GCF_001697945	Neisseria meningitidis	M25476	19400	GCA_001697945.1	2.16819	Complete Genome	11	2	3	4	3	8	4	6
915	GCF_001697965	Neisseria meningitidis	M25456	19400	GCA_001697965.1	2.17311	Complete Genome	11	2	3	4	3	8	4	6
916	GCF_001697985	Neisseria meningitidis	M25419	19400	GCA_001697985.1	2.18956	Complete Genome	11	2	3	4	3	8	4	6
917	GCF_001698005	Neisseria meningitidis	M22740	19400	GCA_001698005.1	2.1729	Complete Genome	11	2	3	4	3	8	4	6
918	GCF_001698025	Neisseria meningitidis	M22822	19400	GCA_001698025.1	2.1739	Complete Genome	2881	179	7	4	56	26	18	8
919	GCF_001698045	Neisseria meningitidis	M07162	19400	GCA_001698045.1	2.19374	Complete Genome	11	2	3	4	3	8	4	6
920	GCF_001698065	Neisseria meningitidis	M08000	19400	GCA_001698065.1	2.16238	Complete Genome	11	2	3	4	3	8	4	6
921	GCF_001698085	Neisseria meningitidis	M24705	19400	GCA_001698085.1	2.17583	Complete Genome	8638	179	7	4	56	26	18	12
922	GCF_001698105	Neisseria meningitidis	M22191	19400	GCA_001698105.1	2.15549	Complete Genome	11	2	3	4	3	8	4	6
923	GCF_001703675	Neisseria meningitidis	M07165	19400	GCA_001703675.1	2.20717	Chromosome	11	2	3	4	3	8	4	6
924	GCF_001703695	Neisseria meningitidis	M22747	19400	GCA_001703695.1	2.17301	Chromosome	11	2	3	4	3	8	4	6

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
925	GCF_001703715	Neisseria meningitidis	M07999	19400	GCA_001703715.1	2.16208	Chromosome	11	2	3	4	3	8	4	6
926	GCF_001703755	Neisseria meningitidis	M22797	19400	GCA_001703755.1	2.1935	Chromosome	11	2	3	4	3	8	4	6
927	GCF_001703795	Neisseria meningitidis	M25073	19400	GCA_001703795.1	2.2044	Chromosome	11	2	3	4	3	8	4	6
928	GCF_001721325	Neisseria meningitidis	VB13856	19400	GCA_001721325.1	2.11465	Contig	6928	222	3	58	7	30	5	255
929	GCF_001721335	Neisseria meningitidis	VB15548	19400	GCA_001721335.1	2.11969	Contig	6928	222	3	58	7	30	5	255
930	GCF_001742295	Neisseria meningitidis	VB493	19400	GCA_001742295.1	2.10228	Contig	6928	222	3	58	7	30	5	255
931	GCF_900017035	Neisseria meningitidis	2842STDY5881246	19400	GCA_900017035.1	2.16638	Scaffold	3720	8	10	313	4	5	3	8
932	GCF_900017045	Neisseria meningitidis	2842STDY5881282	19400	GCA_900017045.1	2.20166	Scaffold	213	7	5	1	13	36	53	15
933	GCF_900017055	Neisseria meningitidis	2842STDY5881285	19400	GCA_900017055.1	2.1579	Scaffold	40	3	6	9	5	9	22	9
934	GCF_900017065	Neisseria meningitidis	2842STDY5881315	19400	GCA_900017065.1	2.124	Scaffold	1466	6	5	173	13	5	24	17
935	GCF_900017075	Neisseria meningitidis	2842STDY5881677	19400	GCA_900017075.1	2.1585	Scaffold	32	4	10	5	4	6	3	8
936	GCF_900017085	Neisseria meningitidis	2842STDY5881734	19400	GCA_900017085.1	2.19614	Scaffold	146	8	6	9	9	9	6	9
937	GCF_900017095	Neisseria meningitidis	2842STDY5881361	19400	GCA_900017095.1	2.11243	Scaffold	2680	2	3	7	217	8	5	2
938	GCF_900017215	Neisseria meningitidis	2842STDY5881451	19400	GCA_900017215.1	2.18246	Scaffold	6288	11	5	18	8	442	24	21
939	GCF_900017225	Neisseria meningitidis	2842STDY5881526	19400	GCA_900017225.1	2.17182	Scaffold	41	3	6	9	5	9	6	9
940	GCF_900017465	Neisseria meningitidis	2842STDY5881240	19400	GCA_900017465.1	2.12229	Scaffold	1372	2	3	7	2	8	6	2
941	GCF_900017935	Neisseria meningitidis	2842STDY5881150	19400	GCA_900017935.1	2.12282	Scaffold	4283	2	212	4	3	8	4	6
942	GCF_900017945	Neisseria meningitidis	2842STDY5881410	19400	GCA_900017945.1	2.16181	Scaffold	42	10	6	9	5	9	6	9
943	GCF_900019005	Neisseria meningitidis	2842STDY5881072	19400	GCA_900019005.1	2.16354	Scaffold	3488	3	6	9	5	9	22	250
944	GCF_900019015	Neisseria meningitidis	2842STDY5881308	19400	GCA_900019015.1	2.16655	Scaffold	33	8	10	5	4	6	3	8
945	GCF_900019895	Neisseria meningitidis	2842STDY5880978	19400	GCA_900019895.1	2.10602	Scaffold	51	2	3	4	23	8	6	6
946	GCF_900019905	Neisseria meningitidis	2842STDY5881159	19400	GCA_900019905.1	2.17084	Scaffold	33	8	10	5	4	6	3	8
947	GCF_900019915	Neisseria meningitidis	2842STDY5881288	19400	GCA_900019915.1	2.16861	Scaffold	2925	3	6	9	5	252	6	9
948	GCF_900019925	Neisseria meningitidis	2842STDY5881293	19400	GCA_900019925.1	2.09208	Scaffold	23	10	5	18	9	11	9	17
949	GCF_900020185	Neisseria meningitidis	2842STDY5881412	19400	GCA_900020185.1	2.19789	Scaffold	41	3	6	9	5	9	6	9
950	GCF_900021115	Neisseria meningitidis	2842STDY5881600	19400	GCA_900021115.1	2.16308	Scaffold	41	3	6	9	5	9	6	9
951	GCF_900021125	Neisseria meningitidis	2842STDY5881609	19400	GCA_900021125.1	2.1262	Scaffold	11	2	3	4	3	8	4	6
952	GCF_900021135	Neisseria meningitidis	2842STDY5881682	19400	GCA_900021135.1	2.20702	Scaffold	41	3	6	9	5	9	6	9
953	GCF_900021945	Neisseria meningitidis	2842STDY5881550	19400	GCA_900021945.1	2.13394	Scaffold	8	2	3	7	2	8	5	2
954	GCF_900021955	Neisseria meningitidis	2842STDY5881590	19400	GCA_900021955.1	2.16894	Scaffold	259	4	10	5	40	6	3	8
955	GCF_900021965	Neisseria meningitidis	2842STDY5881597	19400	GCA_900021965.1	2.11609	Scaffold	32	4	10	5	4	6	3	8
956	GCF_900021985	Neisseria meningitidis	2842STDY5881183	19400	GCA_900021985.1	2.22037	Scaffold	3608	3	116	9	5	5	22	9
957	GCF_900021995	Neisseria meningitidis	2842STDY5881262	19400	GCA_900021995.1	2.1725	Scaffold	7883	10	6	312	5	9	3	9

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
958	GCF_900022005	Neisseria meningitidis	2842STDY5881277	19400	GCA_900022005.1	2.11837	Scaffold	32	4	10	5	4	6	3	8
959	GCF_900022015	Neisseria meningitidis	2842STDY5881380	19400	GCA_900022015.1	2.10389	Scaffold	2631	3	6	9	9	9	6	206
960	GCF_900022025	Neisseria meningitidis	2842STDY5881461	19400	GCA_900022025.1	2.17907	Scaffold	213	7	5	1	13	36	53	15
961	GCF_900022035	Neisseria meningitidis	2842STDY5881479	19400	GCA_900022035.1	2.16556	Scaffold	5266	134	10	15	9	8	11	9
962	GCF_900022045	Neisseria meningitidis	2842STDY5881496	19400	GCA_900022045.1	2.1592	Scaffold	154	3	6	9	5	11	6	9
963	GCF_900022335	Neisseria meningitidis	2842STDY5880977	19400	GCA_900022335.1	2.20242	Scaffold	2286	3	6	9	5	58	22	9
964	GCF_900022355	Neisseria meningitidis	2842STDY5881067	19400	GCA_900022355.1	2.19958	Scaffold	33	8	10	5	4	6	3	8
965	GCF_900022365	Neisseria meningitidis	2842STDY5881081	19400	GCA_900022365.1	2.19612	Scaffold	41	3	6	9	5	9	6	9
966	GCF_900022375	Neisseria meningitidis	2842STDY5881099	19400	GCA_900022375.1	2.12229	Scaffold	11	2	3	4	3	8	4	6
967	GCF_900022385	Neisseria meningitidis	2842STDY5881227	19400	GCA_900022385.1	2.192	Scaffold	1255	3	6	9	5	9	6	8
968	GCF_900022395	Neisseria meningitidis	2842STDY5881278	19400	GCA_900022395.1	2.15912	Scaffold	34	8	10	5	4	5	3	8
969	GCF_900022405	Neisseria meningitidis	2842STDY5881693	19400	GCA_900022405.1	2.15285	Scaffold	41	3	6	9	5	9	6	9
970	GCF_900022415	Neisseria meningitidis	2842STDY5881370	19400	GCA_900022415.1	2.12643	Scaffold	11	2	3	4	3	8	4	6
971	GCF_900022425	Neisseria meningitidis	2842STDY5881376	19400	GCA_900022425.1	2.16669	Scaffold	40	3	6	9	5	9	22	9
972	GCF_900022435	Neisseria meningitidis	2842STDY5881390	19400	GCA_900022435.1	2.16256	Scaffold	34	8	10	5	4	5	3	8
973	GCF_900022445	Neisseria meningitidis	2842STDY5881419	19400	GCA_900022445.1	2.12643	Scaffold	4283	2	212	4	3	8	4	6
974	GCF_900022455	Neisseria meningitidis	2842STDY5881462	19400	GCA_900022455.1	2.20695	Scaffold	3720	8	10	313	4	5	3	8
975	GCF_900024325	Neisseria meningitidis	2842STDY5881744	19400	GCA_900024325.1	2.19948	Scaffold	2016	3	116	9	5	9	22	2
976	GCF_900024335	Neisseria meningitidis	2842STDY5881026	19400	GCA_900024335.1	2.21382	Scaffold	42	10	6	9	5	9	6	9
977	GCF_900024345	Neisseria meningitidis	2842STDY5881101	19400	GCA_900024345.1	2.20165	Scaffold	41	3	6	9	5	9	6	9
978	GCF_900024355	Neisseria meningitidis	2842STDY5881209	19400	GCA_900024355.1	2.1587	Scaffold	41	3	6	9	5	9	6	9
979	GCF_900024365	Neisseria meningitidis	2842STDY5881244	19400	GCA_900024365.1	2.16973	Scaffold	259	4	10	5	40	6	3	8
980	GCF_900024375	Neisseria meningitidis	2842STDY5881321	19400	GCA_900024375.1	2.19459	Scaffold	1158	11	5	18	15	11	24	21
981	GCF_900024385	Neisseria meningitidis	2842STDY5881335	19400	GCA_900024385.1	2.16218	Scaffold	2203	10	6	9	5	11	6	9
982	GCF_900024395	Neisseria meningitidis	2842STDY5881551	19400	GCA_900024395.1	2.13048	Scaffold	8	2	3	7	2	8	5	2
983	GCF_900024405	Neisseria meningitidis	2842STDY5881665	19400	GCA_900024405.1	2.17694	Scaffold	41	3	6	9	5	9	6	9
984	GCF_900024425	Neisseria meningitidis	2842STDY5881344	19400	GCA_900024425.1	2.10464	Scaffold	51	2	3	4	23	8	6	6
985	GCF_900024435	Neisseria meningitidis	2842STDY5881367	19400	GCA_900024435.1	2.11759	Scaffold	11	2	3	4	3	8	4	6
986	GCF_900024445	Neisseria meningitidis	2842STDY5881459	19400	GCA_900024445.1	2.12067	Scaffold	11	2	3	4	3	8	4	6
987	GCF_900024925	Neisseria meningitidis	2842STDY5881474	19400	GCA_900024925.1	2.15838	Scaffold	2925	3	6	9	5	252	6	9
988	GCF_900025435	Neisseria meningitidis	2842STDY5881574	19400	GCA_900025435.1	2.12789	Scaffold	8	2	3	7	2	8	5	2
989	GCF_900025445	Neisseria meningitidis	2842STDY5881587	19400	GCA_900025445.1	2.20552	Scaffold	41	3	6	9	5	9	6	9
990	GCF_900025455	Neisseria meningitidis	2842STDY5881625	19400	GCA_900025455.1	2.16322	Scaffold	41	3	6	9	5	9	6	9
991	GCF_900025465	Neisseria meningitidis	2842STDY5881714	19400	GCA_900025465.1	2.16563	Scaffold	40	3	6	9	5	9	22	9
992	GCF_900025475	Neisseria meningitidis	2842STDY5881692	19400	GCA_900025475.1	2.1513	Scaffold	41	3	6	9	5	9	6	9

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
993	GCF_900025485	Neisseria meningitidis	2842STDY5881050	19400	GCA_900025485.1	2.16423	Scaffold	2708	3	6	9	5	9	201	9
994	GCF_900025505	Neisseria meningitidis	2842STDY5881087	19400	GCA_900025505.1	2.10872	Scaffold	2680	2	3	7	217	8	5	2
995	GCF_900025515	Neisseria meningitidis	2842STDY5881092	19400	GCA_900025515.1	2.17388	Scaffold	2916	10	6	9	5	9	6	12
996	GCF_900025525	Neisseria meningitidis	2842STDY5881156	19400	GCA_900025525.1	2.18904	Scaffold	3553	4	10	45	17	8	11	8
997	GCF_900025535	Neisseria meningitidis	2842STDY5881198	19400	GCA_900025535.1	2.19348	Scaffold	5458	4	6	9	9	58	6	16
998	GCF_900025545	Neisseria meningitidis	2842STDY5881225	19400	GCA_900025545.1	2.17218	Scaffold	3720	8	10	313	4	5	3	8
999	GCF_900025555	Neisseria meningitidis	2842STDY5881298	19400	GCA_900025555.1	2.16784	Scaffold	3720	8	10	313	4	5	3	8
1000	GCF_900025565	Neisseria meningitidis	2842STDY5881305	19400	GCA_900025565.1	2.16837	Scaffold	5266	134	10	15	9	8	11	9
1001	GCF_900025575	Neisseria meningitidis	2842STDY5881505	19400	GCA_900025575.1	2.11002	Scaffold	8	2	3	7	2	8	5	2
1002	GCF_900025585	Neisseria meningitidis	2842STDY5881530	19400	GCA_900025585.1	2.11778	Scaffold	1287	2	3	4	17	8	4	6
1003	GCF_900025595	Neisseria meningitidis	2842STDY5881585	19400	GCA_900025595.1	2.16209	Scaffold	41	3	6	9	5	9	6	9
1004	GCF_900025605	Neisseria meningitidis	2842STDY5881610	19400	GCA_900025605.1	2.1293	Scaffold	11	2	3	4	3	8	4	6
1005	GCF_900025615	Neisseria meningitidis	2842STDY5881626	19400	GCA_900025615.1	2.16271	Scaffold	41	3	6	9	5	9	6	9
1006	GCF_900025625	Neisseria meningitidis	2842STDY5881742	19400	GCA_900025625.1	2.21259	Scaffold	41	3	6	9	5	9	6	9
1007	GCF_900025635	Neisseria meningitidis	2842STDY5881737	19400	GCA_900025635.1	2.15013	Scaffold	482	3	6	9	17	9	6	9
1008	GCF_900025645	Neisseria meningitidis	2842STDY5881343	19400	GCA_900025645.1	2.19777	Scaffold	2286	3	6	9	5	58	22	9
1009	GCF_900025655	Neisseria meningitidis	2842STDY5881347	19400	GCA_900025655.1	2.14805	Scaffold	41	3	6	9	5	9	6	9
1010	GCF_900025665	Neisseria meningitidis	2842STDY5881366	19400	GCA_900025665.1	2.15202	Scaffold	40	3	6	9	5	9	22	9
1011	GCF_900025675	Neisseria meningitidis	2842STDY5881393	19400	GCA_900025675.1	2.15817	Scaffold	41	3	6	9	5	9	6	9
1012	GCF_900025685	Neisseria meningitidis	2842STDY5881406	19400	GCA_900025685.1	2.10687	Scaffold	2699	2	3	7	24	8	5	2
1013	GCF_900025695	Neisseria meningitidis	2842STDY5881415	19400	GCA_900025695.1	2.13511	Scaffold	3552	4	10	5	4	15	3	8
1014	GCF_900025705	Neisseria meningitidis	2842STDY5881414	19400	GCA_900025705.1	2.11608	Scaffold	11	2	3	4	3	8	4	6
1015	GCF_900025715	Neisseria meningitidis	2842STDY5881427	19400	GCA_900025715.1	2.12263	Scaffold	11	2	3	4	3	8	4	6
1016	GCF_900026725	Neisseria meningitidis	2842STDY5881557	19400	GCA_900026725.1	2.1706	Scaffold	41	3	6	9	5	9	6	9
1017	GCF_900026735	Neisseria meningitidis	2842STDY5881664	19400	GCA_900026735.1	2.17916	Scaffold	41	3	6	9	5	9	6	9
1018	GCF_900026745	Neisseria meningitidis	2842STDY5880999	19400	GCA_900026745.1	2.09784	Scaffold	5450	2	3	4	3	8	4	8
1019	GCF_900026755	Neisseria meningitidis	2842STDY5881001	19400	GCA_900026755.1	2.16593	Scaffold	41	3	6	9	5	9	6	9
1020	GCF_900026765	Neisseria meningitidis	2842STDY5881105	19400	GCA_900026765.1	2.18938	Scaffold	42	10	6	9	5	9	6	9
1021	GCF_900026775	Neisseria meningitidis	2842STDY5881161	19400	GCA_900026775.1	2.15932	Scaffold	34	8	10	5	4	5	3	8
1022	GCF_900026785	Neisseria meningitidis	2842STDY5881162	19400	GCA_900026785.1	2.11918	Scaffold	11	2	3	4	3	8	4	6
1023	GCF_900026795	Neisseria meningitidis	2842STDY5881235	19400	GCA_900026795.1	2.15126	Scaffold	1374	3	7	9	5	9	22	9
1024	GCF_900026805	Neisseria meningitidis	2842STDY5881273	19400	GCA_900026805.1	2.16516	Scaffold	3720	8	10	313	4	5	3	8
1025	GCF_900026815	Neisseria meningitidis	2842STDY5881362	19400	GCA_900026815.1	2.15716	Scaffold	42	10	6	9	5	9	6	9
1026	GCF_900026825	Neisseria meningitidis	2842STDY5881388	19400	GCA_900026825.1	2.12723	Scaffold	11	2	3	4	3	8	4	6
1027	GCF_900026835	Neisseria meningitidis	2842STDY5881389	19400	GCA_900026835.1	2.18897	Scaffold	41	3	6	9	5	9	6	9

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1028	GCF_900026845	Neisseria meningitidis	2842STDY5881401	19400	GCA_900026845.1	2.20259	Scaffold	41	3	6	9	5	9	6	9
1029	GCF_900026855	Neisseria meningitidis	2842STDY5881422	19400	GCA_900026855.1	2.1933	Scaffold	3553	4	10	45	17	8	11	8
1030	GCF_900026875	Neisseria meningitidis	2842STDY5881446	19400	GCA_900026875.1	2.11263	Scaffold	167	2	7	6	17	16	18	8
1031	GCF_900026885	Neisseria meningitidis	2842STDY5881509	19400	GCA_900026885.1	2.13448	Scaffold	2174	2	3	7	2	8	15	2
1032	GCF_900027405	Neisseria meningitidis	2842STDY5880970	19400	GCA_900027405.1	2.24852	Scaffold	5445	3	6	9	5	9	22	2
1033	GCF_900027415	Neisseria meningitidis	2842STDY5881632	19400	GCA_900027415.1	2.16351	Scaffold	34	8	10	5	4	5	3	8
1034	GCF_900027425	Neisseria meningitidis	2842STDY5881038	19400	GCA_900027425.1	2.12115	Scaffold	11	2	3	4	3	8	4	6
1035	GCF_900027435	Neisseria meningitidis	2842STDY5881108	19400	GCA_900027435.1	2.16367	Scaffold	2696	8	10	5	4	5	3	15
1036	GCF_900027445	Neisseria meningitidis	2842STDY5881181	19400	GCA_900027445.1	2.12868	Scaffold	11	2	3	4	3	8	4	6
1037	GCF_900027455	Neisseria meningitidis	2842STDY5881508	19400	GCA_900027455.1	2.13804	Scaffold	2174	2	3	7	2	8	15	2
1038	GCF_900027465	Neisseria meningitidis	2842STDY5881518	19400	GCA_900027465.1	2.106	Scaffold	23	10	5	18	9	11	9	17
1039	GCF_900027475	Neisseria meningitidis	2842STDY5881601	19400	GCA_900027475.1	2.16316	Scaffold	41	3	6	9	5	9	6	9
1040	GCF_900027825	Neisseria meningitidis	2842STDY5881010	19400	GCA_900027825.1	2.15298	Contig	34	8	10	5	4	5	3	8
1041	GCF_900027835	Neisseria meningitidis	2842STDY5881036	19400	GCA_900027835.1	2.12431	Scaffold	11	2	3	4	3	8	4	6
1042	GCF_900027845	Neisseria meningitidis	2842STDY5881047	19400	GCA_900027845.1	2.17559	Scaffold	269	4	10	15	9	8	11	9
1043	GCF_900027855	Neisseria meningitidis	2842STDY5881061	19400	GCA_900027855.1	2.17622	Scaffold	269	4	10	15	9	8	11	9
1044	GCF_900027865	Neisseria meningitidis	2842STDY5881089	19400	GCA_900027865.1	2.16552	Scaffold	41	3	6	9	5	9	6	9
1045	GCF_900027875	Neisseria meningitidis	2842STDY5881147	19400	GCA_900027875.1	2.16802	Scaffold	33	8	10	5	4	6	3	8
1046	GCF_900027885	Neisseria meningitidis	2842STDY5881200	19400	GCA_900027885.1	2.10643	Scaffold	2680	2	3	7	217	8	5	2
1047	GCF_900027895	Neisseria meningitidis	2842STDY5881269	19400	GCA_900027895.1	2.27115	Scaffold	41	3	6	9	5	9	6	9
1048	GCF_900027905	Neisseria meningitidis	2842STDY5881525	19400	GCA_900027905.1	2.16242	Scaffold	41	3	6	9	5	9	6	9
1049	GCF_900027915	Neisseria meningitidis	2842STDY5881333	19400	GCA_900027915.1	2.16064	Scaffold	41	3	6	9	5	9	6	9
1050	GCF_900027925	Neisseria meningitidis	2842STDY5881336	19400	GCA_900027925.1	2.14402	Scaffold	41	3	6	9	5	9	6	9
1051	GCF_900027935	Neisseria meningitidis	2842STDY5881598	19400	GCA_900027935.1	2.11977	Scaffold	32	4	10	5	4	6	3	8
1052	GCF_900027945	Neisseria meningitidis	2842STDY5881683	19400	GCA_900027945.1	2.20751	Scaffold	41	3	6	9	5	9	6	9
1053	GCF_900027955	Neisseria meningitidis	2842STDY5881718	19400	GCA_900027955.1	2.15465	Scaffold	41	3	6	9	5	9	6	9
1054	GCF_900027965	Neisseria meningitidis	2842STDY5881372	19400	GCA_900027965.1	2.17902	Scaffold	269	4	10	15	9	8	11	9
1055	GCF_900027975	Neisseria meningitidis	2842STDY5881385	19400	GCA_900027975.1	2.11784	Scaffold	11	2	3	4	3	8	4	6
1056	GCF_900027985	Neisseria meningitidis	2842STDY5881416	19400	GCA_900027985.1	2.12824	Scaffold	11	2	3	4	3	8	4	6
1057	GCF_900027995	Neisseria meningitidis	2842STDY5881431	19400	GCA_900027995.1	2.12082	Scaffold	11	2	3	4	3	8	4	6
1058	GCF_900028005	Neisseria meningitidis	2842STDY5881435	19400	GCA_900028005.1	2.12485	Scaffold	11	2	3	4	3	8	4	6
1059	GCF_900028015	Neisseria meningitidis	2842STDY5881456	19400	GCA_900028015.1	2.15919	Scaffold	3720	8	10	313	4	5	3	8
1060	GCF_900028025	Neisseria meningitidis	2842STDY5881465	19400	GCA_900028025.1	2.16515	Scaffold	32	4	10	5	4	6	3	8
1061	GCF_900028035	Neisseria meningitidis	2842STDY5881466	19400	GCA_900028035.1	2.21574	Scaffold	213	7	5	1	13	36	53	15
1062	GCF_900028045	Neisseria meningitidis	2842STDY5881468	19400	GCA_900028045.1	2.16693	Scaffold	3720	8	10	313	4	5	3	8

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1063	GCF_900028055	Neisseria meningitidis	2842STDY5881472	19400	GCA_900028055.1	2.14755	Scaffold	40	3	6	9	5	9	22	9
1064	GCF_900028065	Neisseria meningitidis	2842STDY5881473	19400	GCA_900028065.1	2.15826	Scaffold	2925	3	6	9	5	252	6	9
1065	GCF_900028075	Neisseria meningitidis	2842STDY5881481	19400	GCA_900028075.1	2.09688	Scaffold	1655	12	5	18	9	11	9	17
1066	GCF_900028085	Neisseria meningitidis	2842STDY5881483	19400	GCA_900028085.1	2.12992	Scaffold	1466	6	5	173	13	5	24	17
1067	GCF_900028095	Neisseria meningitidis	2842STDY5881485	19400	GCA_900028095.1	2.19601	Scaffold	1158	11	5	18	15	11	24	21
1068	GCF_900028105	Neisseria meningitidis	2842STDY5881487	19400	GCA_900028105.1	2.19962	Scaffold	9808	3	6	654	5	9	22	9
1069	GCF_900028115	Neisseria meningitidis	2842STDY5881493	19400	GCA_900028115.1	2.155	Scaffold	41	3	6	9	5	9	6	9
1070	GCF_900029025	Neisseria meningitidis	2842STDY5881112	19400	GCA_900029025.1	2.11178	Scaffold	2699	2	3	7	24	8	5	2
1071	GCF_900029035	Neisseria meningitidis	2842STDY5881254	19400	GCA_900029035.1	2.19874	Scaffold	213	7	5	1	13	36	53	15
1072	GCF_900029045	Neisseria meningitidis	2842STDY5881621	19400	GCA_900029045.1	2.16552	Scaffold	41	3	6	9	5	9	6	9
1073	GCF_900029055	Neisseria meningitidis	2842STDY5881349	19400	GCA_900029055.1	2.11316	Scaffold	11	2	3	4	3	8	4	6
1074	GCF_900029065	Neisseria meningitidis	2842STDY5881352	19400	GCA_900029065.1	2.09713	Scaffold	11	2	3	4	3	8	4	6
1075	GCF_900029075	Neisseria meningitidis	2842STDY5881373	19400	GCA_900029075.1	2.12891	Scaffold	11	2	3	4	3	8	4	6
1076	GCF_900029085	Neisseria meningitidis	2842STDY5881381	19400	GCA_900029085.1	2.18004	Scaffold	269	4	10	15	9	8	11	9
1077	GCF_900029095	Neisseria meningitidis	2842STDY5881413	19400	GCA_900029095.1	2.1224	Scaffold	11	2	3	4	3	8	4	6
1078	GCF_900029105	Neisseria meningitidis	2842STDY5881428	19400	GCA_900029105.1	2.12953	Scaffold	11	2	3	4	3	8	4	6
1079	GCF_900029115	Neisseria meningitidis	2842STDY5881469	19400	GCA_900029115.1	2.11736	Scaffold	32	4	10	5	4	6	3	8
1080	GCF_900029225	Neisseria meningitidis	2842STDY5881717	19400	GCA_900029225.1	2.15725	Scaffold	41	3	6	9	5	9	6	9
1081	GCF_900029235	Neisseria meningitidis	2842STDY5881111	19400	GCA_900029235.1	2.20199	Scaffold	42	10	6	9	5	9	6	9
1082	GCF_900029245	Neisseria meningitidis	2842STDY5881317	19400	GCA_900029245.1	2.15666	Scaffold	34	8	10	5	4	5	3	8
1083	GCF_900029255	Neisseria meningitidis	2842STDY5881563	19400	GCA_900029255.1	2.24317	Scaffold	44	9	6	9	9	9	6	9
1084	GCF_900029265	Neisseria meningitidis	2842STDY5881377	19400	GCA_900029265.1	2.11445	Scaffold	32	4	10	5	4	6	3	8
1085	GCF_900029275	Neisseria meningitidis	2842STDY5881363	19400	GCA_900029275.1	2.22492	Scaffold	42	10	6	9	5	9	6	9
1086	GCF_900029805	Neisseria meningitidis	2842STDY5881723	19400	GCA_900029805.1	2.16315	Scaffold	41	3	6	9	5	9	6	9
1087	GCF_900029825	Neisseria meningitidis	2842STDY5881337	19400	GCA_900029825.1	2.16839	Scaffold	42	10	6	9	5	9	6	9
1088	GCF_900029835	Neisseria meningitidis	2842STDY5881630	19400	GCA_900029835.1	2.11499	Scaffold	153	2	3	7	2	34	5	2
1089	GCF_900029845	Neisseria meningitidis	2842STDY5881633	19400	GCA_900029845.1	2.16107	Scaffold	34	8	10	5	4	5	3	8
1090	GCF_900029855	Neisseria meningitidis	2842STDY5881382	19400	GCA_900029855.1	2.16045	Scaffold	41	3	6	9	5	9	6	9
1091	GCF_900030875	Neisseria meningitidis	2842STDY5880972	19400	GCA_900030875.1	2.15655	Scaffold	41	3	6	9	5	9	6	9
1092	GCF_900030885	Neisseria meningitidis	2842STDY5881733	19400	GCA_900030885.1	2.20036	Scaffold	146	8	6	9	9	9	6	9
1093	GCF_900030895	Neisseria meningitidis	2842STDY5880989	19400	GCA_900030895.1	2.15179	Scaffold	41	3	6	9	5	9	6	9
1094	GCF_900030905	Neisseria meningitidis	2842STDY5880995	19400	GCA_900030905.1	2.11162	Scaffold	11	2	3	4	3	8	4	6
1095	GCF_900030915	Neisseria meningitidis	2842STDY5881015	19400	GCA_900030915.1	2.11217	Scaffold	2680	2	3	7	217	8	5	2
1096	GCF_900030925	Neisseria meningitidis	2842STDY5881018	19400	GCA_900030925.1	2.1449	Scaffold	42	10	6	9	5	9	6	9
1097	GCF_900030935	Neisseria meningitidis	2842STDY5881051	19400	GCA_900030935.1	2.14633	Scaffold	41	3	6	9	5	9	6	9

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1098	GCF_900030945	Neisseria meningitidis	2842STDY5881054	19400	GCA_900030945.1	2.11551	Scaffold	32	4	10	5	4	6	3	8
1099	GCF_900030955	Neisseria meningitidis	2842STDY5881056	19400	GCA_900030955.1	2.21977	Scaffold	461	12	5	12	35	60	22	17
1100	GCF_900030965	Neisseria meningitidis	2842STDY5881073	19400	GCA_900030965.1	2.16001	Scaffold	41	3	6	9	5	9	6	9
1101	GCF_900030975	Neisseria meningitidis	2842STDY5881093	19400	GCA_900030975.1	2.19524	Scaffold	1163	4	10	2	5	3	11	9
1102	GCF_900030985	Neisseria meningitidis	2842STDY5881115	19400	GCA_900030985.1	2.114	Scaffold	2680	2	3	7	217	8	5	2
1103	GCF_900030995	Neisseria meningitidis	2842STDY5881120	19400	GCA_900030995.1	2.14724	Scaffold	41	3	6	9	5	9	6	9
1104	GCF_900031005	Neisseria meningitidis	2842STDY5881126	19400	GCA_900031005.1	2.16405	Scaffold	42	10	6	9	5	9	6	9
1105	GCF_900031015	Neisseria meningitidis	2842STDY5881131	19400	GCA_900031015.1	2.19482	Scaffold	3549	132	7	11	17	62	21	2
1106	GCF_900031025	Neisseria meningitidis	2842STDY5881154	19400	GCA_900031025.1	2.12759	Scaffold	11	2	3	4	3	8	4	6
1107	GCF_900031035	Neisseria meningitidis	2842STDY5881190	19400	GCA_900031035.1	2.13019	Scaffold	11	2	3	4	3	8	4	6
1108	GCF_900031045	Neisseria meningitidis	2842STDY5881199	19400	GCA_900031045.1	2.11669	Scaffold	8	2	3	7	2	8	5	2
1109	GCF_900031055	Neisseria meningitidis	2842STDY5881223	19400	GCA_900031055.1	2.10752	Scaffold	167	2	7	6	17	16	18	8
1110	GCF_900031065	Neisseria meningitidis	2842STDY5881241	19400	GCA_900031065.1	2.16727	Scaffold	213	7	5	1	13	36	53	15
1111	GCF_900031385	Neisseria meningitidis	2842STDY5880996	19400	GCA_900031385.1	2.18082	Scaffold	1286	11	5	168	8	11	4	21
1112	GCF_900031395	Neisseria meningitidis	2842STDY5881079	19400	GCA_900031395.1	2.12069	Scaffold	11	2	3	4	3	8	4	6
1113	GCF_900031405	Neisseria meningitidis	2842STDY5881113	19400	GCA_900031405.1	2.11901	Scaffold	11	2	3	4	3	8	4	6
1114	GCF_900031425	Neisseria meningitidis	2842STDY5881249	19400	GCA_900031425.1	2.12846	Scaffold	11	2	3	4	3	8	4	6
1115	GCF_900031435	Neisseria meningitidis	2842STDY5881309	19400	GCA_900031435.1	2.09838	Scaffold	1655	12	5	18	9	11	9	17
1116	GCF_900031445	Neisseria meningitidis	2842STDY5881310	19400	GCA_900031445.1	2.0951	Scaffold	41	3	6	9	5	9	6	9
1117	GCF_900031455	Neisseria meningitidis	2842STDY5881724	19400	GCA_900031455.1	2.15678	Scaffold	41	3	6	9	5	9	6	9
1118	GCF_900031465	Neisseria meningitidis	2842STDY5881346	19400	GCA_900031465.1	2.1533	Scaffold	41	3	6	9	5	9	6	9
1119	GCF_900031475	Neisseria meningitidis	2842STDY5881450	19400	GCA_900031475.1	2.13316	Scaffold	1466	6	5	173	13	5	24	17
1120	GCF_900031545	Neisseria meningitidis	2842STDY5881588	19400	GCA_900031545.1	2.20875	Scaffold	41	3	6	9	5	9	6	9
1121	GCF_900032015	Neisseria meningitidis	2842STDY5880958	19400	GCA_900032015.1	2.16366	Scaffold	41	3	6	9	5	9	6	9
1122	GCF_900032025	Neisseria meningitidis	2842STDY5881012	19400	GCA_900032025.1	2.13357	Scaffold	11	2	3	4	3	8	4	6
1123	GCF_900032035	Neisseria meningitidis	2842STDY5881013	19400	GCA_900032035.1	2.19794	Scaffold	1403	10	6	63	5	9	6	12
1124	GCF_900032045	Neisseria meningitidis	2842STDY5881572	19400	GCA_900032045.1	2.16215	Scaffold	41	3	6	9	5	9	6	9
1125	GCF_900032055	Neisseria meningitidis	2842STDY5881342	19400	GCA_900032055.1	2.17575	Scaffold	269	4	10	15	9	8	11	9
1126	GCF_900032065	Neisseria meningitidis	2842STDY5881355	19400	GCA_900032065.1	2.15773	Scaffold	41	3	6	9	5	9	6	9
1127	GCF_900032075	Neisseria meningitidis	2842STDY5881398	19400	GCA_900032075.1	2.11973	Scaffold	11	2	3	4	3	8	4	6
1128	GCF_900032085	Neisseria meningitidis	2842STDY5881423	19400	GCA_900032085.1	2.16791	Scaffold	33	8	10	5	4	6	3	8
1129	GCF_900032095	Neisseria meningitidis	2842STDY5881463	19400	GCA_900032095.1	2.15189	Scaffold	1374	3	7	9	5	9	22	9
1130	GCF_900032105	Neisseria meningitidis	2842STDY5881482	19400	GCA_900032105.1	2.10042	Scaffold	41	3	6	9	5	9	6	9
1131	GCF_900032415	Neisseria meningitidis	2842STDY5881687	19400	GCA_900032415.1	2.16652	Scaffold	41	3	6	9	5	9	6	9
1132	GCF_900032425	Neisseria meningitidis	2842STDY5881751	19400	GCA_900032425.1	2.1666	Scaffold	259	4	10	5	40	6	3	8

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1133	GCF_900032435	Neisseria meningitidis	2842STDY5880997	19400	GCA_900032435.1	2.16391	Contig	259	4	10	5	40	6	3	8
1134	GCF_900032445	Neisseria meningitidis	2842STDY5881014	19400	GCA_900032445.1	2.17402	Scaffold	283	4	10	15	9	8	11	17
1135	GCF_900032455	Neisseria meningitidis	2842STDY5881069	19400	GCA_900032455.1	2.12756	Scaffold	11	2	3	4	3	8	4	6
1136	GCF_900032465	Neisseria meningitidis	2842STDY5881086	19400	GCA_900032465.1	2.13097	Scaffold	2704	2	3	4	3	8	214	6
1137	GCF_900032475	Neisseria meningitidis	2842STDY5881096	19400	GCA_900032475.1	2.19834	Scaffold	41	3	6	9	5	9	6	9
1138	GCF_900032485	Neisseria meningitidis	2842STDY5881229	19400	GCA_900032485.1	2.18544	Scaffold	6288	11	5	18	8	442	24	21
1139	GCF_900032495	Neisseria meningitidis	2842STDY5881250	19400	GCA_900032495.1	2.21152	Scaffold	40	3	6	9	5	9	22	9
1140	GCF_900032505	Neisseria meningitidis	2842STDY5881513	19400	GCA_900032505.1	2.1285	Scaffold	8	2	3	7	2	8	5	2
1141	GCF_900032515	Neisseria meningitidis	2842STDY5881591	19400	GCA_900032515.1	2.16955	Scaffold	259	4	10	5	40	6	3	8
1142	GCF_900032525	Neisseria meningitidis	2842STDY5881356	19400	GCA_900032525.1	2.16046	Scaffold	4065	8	10	5	4	5	3	7
1143	GCF_900032535	Neisseria meningitidis	2842STDY5881360	19400	GCA_900032535.1	2.16766	Scaffold	283	4	10	15	9	8	11	17
1144	GCF_900032545	Neisseria meningitidis	2842STDY5881369	19400	GCA_900032545.1	2.19197	Scaffold	146	8	6	9	9	9	6	9
1145	GCF_900032555	Neisseria meningitidis	2842STDY5881392	19400	GCA_900032555.1	2.10848	Scaffold	2680	2	3	7	217	8	5	2
1146	GCF_900032565	Neisseria meningitidis	2842STDY5881402	19400	GCA_900032565.1	2.19351	Scaffold	42	10	6	9	5	9	6	9
1147	GCF_900032575	Neisseria meningitidis	2842STDY5881418	19400	GCA_900032575.1	2.16064	Scaffold	33	8	10	5	4	6	3	8
1148	GCF_900032585	Neisseria meningitidis	2842STDY5881420	19400	GCA_900032585.1	2.12076	Scaffold	11	2	3	4	3	8	4	6
1149	GCF_900032595	Neisseria meningitidis	2842STDY5881425	19400	GCA_900032595.1	2.12831	Scaffold	11	2	3	4	3	8	4	6
1150	GCF_900032605	Neisseria meningitidis	2842STDY5881436	19400	GCA_900032605.1	2.20182	Scaffold	5458	4	6	9	9	58	6	16
1151	GCF_900032615	Neisseria meningitidis	2842STDY5881441	19400	GCA_900032615.1	2.15857	Scaffold	41	3	6	9	5	9	6	9
1152	GCF_900032625	Neisseria meningitidis	2842STDY5881443	19400	GCA_900032625.1	2.15965	Scaffold	34	8	10	5	4	5	3	8
1153	GCF_900032635	Neisseria meningitidis	2842STDY5881464	19400	GCA_900032635.1	2.16841	Scaffold	7883	10	6	312	5	9	3	9
1154	GCF_900032645	Neisseria meningitidis	2842STDY5881486	19400	GCA_900032645.1	2.0892	Scaffold	23	10	5	18	9	11	9	17
1155	GCF_900032655	Neisseria meningitidis	2842STDY5881519	19400	GCA_900032655.1	2.10469	Scaffold	23	10	5	18	9	11	9	17
1156	GCF_900033065	Neisseria meningitidis	2842STDY5881133	19400	GCA_900033065.1	2.1948	Scaffold	41	3	6	9	5	9	6	9
1157	GCF_900033075	Neisseria meningitidis	2842STDY5881168	19400	GCA_900033075.1	2.12278	Scaffold	11	2	3	4	3	8	4	6
1158	GCF_900033085	Neisseria meningitidis	2842STDY5881178	19400	GCA_900033085.1	2.11076	Scaffold	3486	2	3	4	217	8	5	2
1159	GCF_900033095	Neisseria meningitidis	2842STDY5881251	19400	GCA_900033095.1	2.12988	Scaffold	11	2	3	4	3	8	4	6
1160	GCF_900033105	Neisseria meningitidis	2842STDY5881267	19400	GCA_900033105.1	2.16614	Scaffold	32	4	10	5	4	6	3	8
1161	GCF_900033115	Neisseria meningitidis	2842STDY5881268	19400	GCA_900033115.1	2.21532	Scaffold	213	7	5	1	13	36	53	15
1162	GCF_900033125	Neisseria meningitidis	2842STDY5881492	19400	GCA_900033125.1	2.16126	Scaffold	41	3	6	9	5	9	6	9
1163	GCF_900033135	Neisseria meningitidis	2842STDY5881330	19400	GCA_900033135.1	2.19185	Scaffold	9808	3	6	654	5	9	22	9
1164	GCF_900033145	Neisseria meningitidis	2842STDY5881322	19400	GCA_900033145.1	2.08635	Scaffold	23	10	5	18	9	11	9	17
1165	GCF_900033155	Neisseria meningitidis	2842STDY5881497	19400	GCA_900033155.1	2.13563	Scaffold	11	2	3	4	3	8	4	6
1166	GCF_900033165	Neisseria meningitidis	2842STDY5881558	19400	GCA_900033165.1	2.1705	Scaffold	41	3	6	9	5	9	6	9
1167	GCF_900033175	Neisseria meningitidis	2842STDY5881638	19400	GCA_900033175.1	2.12971	Scaffold	11	2	3	4	3	8	4	6

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1168	GCF_900033185	Neisseria meningitidis	2842STDY5881715	19400	GCA_900033185.1	2.17541	Scaffold	40	3	6	9	5	9	22	9
1169	GCF_900033195	Neisseria meningitidis	2842STDY5881354	19400	GCA_900033195.1	2.15556	Scaffold	259	4	10	5	40	6	3	8
1170	GCF_900033205	Neisseria meningitidis	2842STDY5881358	19400	GCA_900033205.1	2.13427	Scaffold	11	2	3	4	3	8	4	6
1171	GCF_900033215	Neisseria meningitidis	2842STDY5881357	19400	GCA_900033215.1	2.16153	Scaffold	34	8	10	5	4	5	3	8
1172	GCF_900033225	Neisseria meningitidis	2842STDY5881368	19400	GCA_900033225.1	2.12408	Scaffold	11	2	3	4	3	8	4	6
1173	GCF_900033235	Neisseria meningitidis	2842STDY5881375	19400	GCA_900033235.1	2.15491	Scaffold	41	3	6	9	5	9	6	9
1174	GCF_900033255	Neisseria meningitidis	2842STDY5881384	19400	GCA_900033255.1	2.2061	Scaffold	33	8	10	5	4	6	3	8
1175	GCF_900033265	Neisseria meningitidis	2842STDY5881399	19400	GCA_900033265.1	2.12698	Scaffold	2704	2	3	4	3	8	214	6
1176	GCF_900033275	Neisseria meningitidis	2842STDY5881403	19400	GCA_900033275.1	2.15877	Scaffold	2696	8	10	5	4	5	3	15
1177	GCF_900033285	Neisseria meningitidis	2842STDY5881405	19400	GCA_900033285.1	2.2085	Scaffold	42	10	6	9	5	9	6	9
1178	GCF_900033295	Neisseria meningitidis	2842STDY5881407	19400	GCA_900033295.1	2.12776	Scaffold	11	2	3	4	3	8	4	6
1179	GCF_900033305	Neisseria meningitidis	2842STDY5881438	19400	GCA_900033305.1	2.11168	Scaffold	8	2	3	7	2	8	5	2
1180	GCF_900033315	Neisseria meningitidis	2842STDY5881447	19400	GCA_900033315.1	2.10666	Scaffold	167	2	7	6	17	16	18	8
1181	GCF_900033325	Neisseria meningitidis	2842STDY5881448	19400	GCA_900033325.1	2.16983	Scaffold	3720	8	10	313	4	5	3	8
1182	GCF_900033335	Neisseria meningitidis	2842STDY5881460	19400	GCA_900033335.1	2.14973	Scaffold	34	8	10	5	4	5	3	8
1183	GCF_900033345	Neisseria meningitidis	2842STDY5881475	19400	GCA_900033345.1	2.15809	Scaffold	2925	3	6	9	5	252	6	9
1184	GCF_900033355	Neisseria meningitidis	2842STDY5881476	19400	GCA_900033355.1	2.07631	Scaffold	23	10	5	18	9	11	9	17
1185	GCF_900033365	Neisseria meningitidis	2842STDY5881514	19400	GCA_900033365.1	2.12859	Scaffold	8	2	3	7	2	8	5	2
1186	GCF_900033615	Neisseria meningitidis	2842STDY5881057	19400	GCA_900033615.1	2.12728	Scaffold	11	2	3	4	3	8	4	6
1187	GCF_900033625	Neisseria meningitidis	2842STDY5881084	19400	GCA_900033625.1	2.1619	Scaffold	34	8	10	5	4	5	3	8
1188	GCF_900033635	Neisseria meningitidis	2842STDY5881253	19400	GCA_900033635.1	2.15701	Scaffold	34	8	10	5	4	5	3	8
1189	GCF_900033645	Neisseria meningitidis	2842STDY5881383	19400	GCA_900033645.1	2.16901	Scaffold	41	3	6	9	5	9	6	9
1190	GCF_900033655	Neisseria meningitidis	2842STDY5881433	19400	GCA_900033655.1	2.12301	Scaffold	11	2	3	4	3	8	4	6
1191	GCF_900033665	Neisseria meningitidis	2842STDY5881470	19400	GCA_900033665.1	2.15827	Scaffold	34	8	10	5	4	5	3	8
1192	GCF_900034195	Neisseria meningitidis	2842STDY5880961	19400	GCA_900034195.1	2.16004	Scaffold	34	8	10	5	4	5	3	8
1193	GCF_900034205	Neisseria meningitidis	2842STDY5881138	19400	GCA_900034205.1	2.12197	Scaffold	11	2	3	4	3	8	4	6
1194	GCF_900034215	Neisseria meningitidis	2842STDY5881142	19400	GCA_900034215.1	2.133	Scaffold	3552	4	10	5	4	15	3	8
1195	GCF_900034225	Neisseria meningitidis	2842STDY5881257	19400	GCA_900034225.1	2.20869	Scaffold	3720	8	10	313	4	5	3	8
1196	GCF_900034235	Neisseria meningitidis	2842STDY5881374	19400	GCA_900034235.1	2.16896	Scaffold	2708	3	6	9	5	9	201	9
1197	GCF_900034245	Neisseria meningitidis	2842STDY5881395	19400	GCA_900034245.1	2.20144	Scaffold	1163	4	10	2	5	3	11	9
1198	GCF_900034255	Neisseria meningitidis	2842STDY5881426	19400	GCA_900034255.1	2.1264	Scaffold	11	2	3	4	3	8	4	6
1199	GCF_900034265	Neisseria meningitidis	2842STDY5881471	19400	GCA_900034265.1	2.19993	Scaffold	213	7	5	1	13	36	53	15
1200	GCF_900034275	Neisseria meningitidis	2842STDY5881477	19400	GCA_900034275.1	2.15578	Scaffold	136	27	6	9	3	9	6	16
1201	GCF_900034525	Neisseria meningitidis	2842STDY5880968	19400	GCA_900034525.1	2.17375	Scaffold	42	10	6	9	5	9	6	9
1202	GCF_900034535	Neisseria meningitidis	2842STDY5881028	19400	GCA_900034535.1	2.16963	Scaffold	42	10	6	9	5	9	6	9

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1203	GCF_900034545	Neisseria meningitidis	2842STDY5881032	19400	GCA_900034545.1	2.1527	Scaffold	40	3	6	9	5	9	22	9
1204	GCF_900034555	Neisseria meningitidis	2842STDY5881033	19400	GCA_900034555.1	2.11303	Contig	11	2	3	4	3	8	4	6
1205	GCF_900034565	Neisseria meningitidis	2842STDY5881134	19400	GCA_900034565.1	2.12902	Scaffold	11	2	3	4	3	8	4	6
1206	GCF_900034575	Neisseria meningitidis	2842STDY5881163	19400	GCA_900034575.1	2.12123	Scaffold	11	2	3	4	3	8	4	6
1207	GCF_900034585	Neisseria meningitidis	2842STDY5881177	19400	GCA_900034585.1	2.12012	Scaffold	11	2	3	4	3	8	4	6
1208	GCF_900034595	Neisseria meningitidis	2842STDY5881187	19400	GCA_900034595.1	2.12303	Scaffold	11	2	3	4	3	8	4	6
1209	GCF_900034605	Neisseria meningitidis	2842STDY5881745	19400	GCA_900034605.1	2.20086	Scaffold	2016	3	116	9	5	9	22	2
1210	GCF_900034615	Neisseria meningitidis	2842STDY5881386	19400	GCA_900034615.1	2.1548	Scaffold	3488	3	6	9	5	9	22	250
1211	GCF_900034625	Neisseria meningitidis	2842STDY5881454	19400	GCA_900034625.1	2.15398	Scaffold	213	7	5	1	13	36	53	15
1212	GCF_900034635	Neisseria meningitidis	2842STDY5881484	19400	GCA_900034635.1	2.15412	Scaffold	34	8	10	5	4	5	3	8
1213	GCF_900035125	Neisseria meningitidis	2842STDY5881144	19400	GCA_900035125.1	2.11916	Scaffold	11	2	3	4	3	8	4	6
1214	GCF_900035135	Neisseria meningitidis	2842STDY5881575	19400	GCA_900035135.1	2.12601	Scaffold	8	2	3	7	2	8	5	2
1215	GCF_900035145	Neisseria meningitidis	2842STDY5881359	19400	GCA_900035145.1	2.20276	Scaffold	1403	10	6	63	5	9	6	12
1216	GCF_900035155	Neisseria meningitidis	2842STDY5881394	19400	GCA_900035155.1	2.16286	Scaffold	2916	10	6	9	5	9	6	12
1217	GCF_900035165	Neisseria meningitidis	2842STDY5881396	19400	GCA_900035165.1	2.15537	Scaffold	41	3	6	9	5	9	6	9
1218	GCF_900035175	Neisseria meningitidis	2842STDY5881430	19400	GCA_900035175.1	2.10724	Scaffold	3486	2	3	4	217	8	5	2
1219	GCF_900035405	Neisseria meningitidis	2842STDY5881720	19400	GCA_900035405.1	2.15953	Scaffold	41	3	6	9	5	9	6	9
1220	GCF_900035415	Neisseria meningitidis	2842STDY5881095	19400	GCA_900035415.1	2.15623	Scaffold	41	3	6	9	5	9	6	9
1221	GCF_900035425	Neisseria meningitidis	2842STDY5881097	19400	GCA_900035425.1	2.13177	Scaffold	11	2	3	4	3	8	4	6
1222	GCF_900035435	Neisseria meningitidis	2842STDY5881495	19400	GCA_900035435.1	2.16161	Scaffold	154	3	6	9	5	11	6	9
1223	GCF_900035445	Neisseria meningitidis	2842STDY5881645	19400	GCA_900035445.1	2.19267	Scaffold	41	3	6	9	5	9	6	9
1224	GCF_900035455	Neisseria meningitidis	2842STDY5881348	19400	GCA_900035455.1	2.16695	Scaffold	4100	3	6	9	15	11	6	9
1225	GCF_900035465	Neisseria meningitidis	2842STDY5881353	19400	GCA_900035465.1	2.09882	Scaffold	5450	2	3	4	3	8	4	8
1226	GCF_900035485	Neisseria meningitidis	2842STDY5881429	19400	GCA_900035485.1	2.12708	Scaffold	11	2	3	4	3	8	4	6
1227	GCF_900035545	Neisseria meningitidis	2842STDY5881644	19400	GCA_900035545.1	2.20372	Scaffold	41	3	6	9	5	9	6	9
1228	GCF_900035555	Neisseria meningitidis	2842STDY5881417	19400	GCA_900035555.1	2.13024	Scaffold	2704	2	3	4	3	8	214	6
1229	GCF_900036235	Neisseria meningitidis	2842STDY5881334	19400	GCA_900036235.1	2.15322	Scaffold	34	8	10	5	4	5	3	8
1230	GCF_900036245	Neisseria meningitidis	2842STDY5881391	19400	GCA_900036245.1	2.12741	Scaffold	2704	2	3	4	3	8	214	6
1231	GCF_900036255	Neisseria meningitidis	2842STDY5881449	19400	GCA_900036255.1	2.19349	Scaffold	1255	3	6	9	5	9	6	8
1232	GCF_900036275	Neisseria meningitidis	2842STDY5881506	19400	GCA_900036275.1	2.12455	Scaffold	8	2	3	7	2	8	5	2
1233	GCF_900036615	Neisseria meningitidis	2842STDY5880973	19400	GCA_900036615.1	2.18189	Scaffold	269	4	10	15	9	8	11	9
1234	GCF_900036695	Neisseria meningitidis	2842STDY5880963	19400	GCA_900036695.1	2.16823	Scaffold	2203	10	6	9	5	11	6	9
1235	GCF_900036705	Neisseria meningitidis	2842STDY5881571	19400	GCA_900036705.1	2.16118	Scaffold	41	3	6	9	5	9	6	9
1236	GCF_900036715	Neisseria meningitidis	2842STDY5881109	19400	GCA_900036715.1	2.16697	Scaffold	40	3	6	9	5	9	22	9
1237	GCF_900036725	Neisseria meningitidis	2842STDY5881287	19400	GCA_900036725.1	2.16591	Scaffold	2925	3	6	9	5	252	6	9

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1238	GCF_900037245	Neisseria meningitidis	2842STDY5881035	19400	GCA_900037245.1	2.19823	Scaffold	146	8	6	9	9	9	6	9
1239	GCF_900037255	Neisseria meningitidis	2842STDY5881146	19400	GCA_900037255.1	2.12778	Scaffold	2704	2	3	4	3	8	214	6
1240	GCF_900037265	Neisseria meningitidis	2842STDY5881197	19400	GCA_900037265.1	2.11719	Scaffold	11	2	3	4	3	8	4	6
1241	GCF_900037275	Neisseria meningitidis	2842STDY5881292	19400	GCA_900037275.1	2.16702	Scaffold	2925	3	6	9	5	252	6	9
1242	GCF_900037285	Neisseria meningitidis	2842STDY5881351	19400	GCA_900037285.1	2.16349	Scaffold	259	4	10	5	40	6	3	8
1243	GCF_900037295	Neisseria meningitidis	2842STDY5881379	19400	GCA_900037295.1	2.12463	Scaffold	11	2	3	4	3	8	4	6
1244	GCF_900037305	Neisseria meningitidis	2842STDY5881408	19400	GCA_900037305.1	2.11515	Scaffold	2680	2	3	7	217	8	5	2
1245	GCF_900037315	Neisseria meningitidis	2842STDY5881458	19400	GCA_900037315.1	2.20714	Scaffold	40	3	6	9	5	9	22	9
1246	GCF_900037505	Neisseria meningitidis	2842STDY5881034	19400	GCA_900037505.1	2.11951	Scaffold	11	2	3	4	3	8	4	6
1247	GCF_900037515	Neisseria meningitidis	2842STDY5881218	19400	GCA_900037515.1	2.16549	Scaffold	42	10	6	9	5	9	6	9
1248	GCF_900037525	Neisseria meningitidis	2842STDY5881688	19400	GCA_900037525.1	2.16848	Scaffold	41	3	6	9	5	9	6	9
1249	GCF_900037535	Neisseria meningitidis	2842STDY5881387	19400	GCA_900037535.1	2.15704	Scaffold	41	3	6	9	5	9	6	9
1250	GCF_900037545	Neisseria meningitidis	2842STDY5881440	19400	GCA_900037545.1	2.17228	Scaffold	22	11	5	18	8	11	24	21
1251	GCF_900037555	Neisseria meningitidis	2842STDY5881498	19400	GCA_900037555.1	2.13666	Scaffold	11	2	3	4	3	8	4	6
1252	GCF_900038035	Neisseria meningitidis	2842STDY5880984	19400	GCA_900038035.1	2.15061	Contig	34	8	10	5	4	5	3	8
1253	GCF_900038045	Neisseria meningitidis	2842STDY5881170	19400	GCA_900038045.1	2.12353	Scaffold	11	2	3	4	3	8	4	6
1254	GCF_900038055	Neisseria meningitidis	2842STDY5881411	19400	GCA_900038055.1	2.20571	Scaffold	3549	132	7	11	17	62	21	2
1255	GCF_900038075	Neisseria meningitidis	2842STDY5880998	19400	GCA_900038075.1	2.10297	Scaffold	11	2	3	4	3	8	4	6
1256	GCF_900038645	Neisseria meningitidis	2842STDY5881066	19400	GCA_900038645.1	2.16972	Scaffold	41	3	6	9	5	9	6	9
1257	GCF_900038655	Neisseria meningitidis	2842STDY5881478	19400	GCA_900038655.1	2.16761	Scaffold	3720	8	10	313	4	5	3	8
1258	GCF_900038675	Neisseria meningitidis	2842STDY5881421	19400	GCA_900038675.1	2.12798	Scaffold	11	2	3	4	3	8	4	6
1259	GCF_900039285	Neisseria meningitidis	2842STDY5881434	19400	GCA_900039285.1	2.1264	Scaffold	11	2	3	4	3	8	4	6
1260	GCF_900039295	Neisseria meningitidis	2842STDY5881480	19400	GCA_900039295.1	2.16624	Scaffold	33	8	10	5	4	6	3	8
1261	GCF_900039415	Neisseria meningitidis	2842STDY5880969	19400	GCA_900039415.1	2.183	Contig	269	4	10	15	9	8	11	9
1262	GCF_900039425	Neisseria meningitidis	2842STDY5881000	19400	GCA_900039425.1	2.15322	Scaffold	259	4	10	5	40	6	3	8
1263	GCF_900039435	Neisseria meningitidis	2842STDY5880992	19400	GCA_900039435.1	2.1646	Scaffold	4100	3	6	9	15	11	6	9
1264	GCF_900039445	Neisseria meningitidis	2842STDY5881153	19400	GCA_900039445.1	2.12878	Scaffold	11	2	3	4	3	8	4	6
1265	GCF_900039585	Neisseria meningitidis	2842STDY5881741	19400	GCA_900039585.1	2.21345	Scaffold	41	3	6	9	5	9	6	9
1266	GCF_900039595	Neisseria meningitidis	2842STDY5881728	19400	GCA_900039595.1	2.16278	Scaffold	42	10	6	9	5	9	6	9
1267	GCF_900039605	Neisseria meningitidis	2842STDY5881457	19400	GCA_900039605.1	2.12115	Scaffold	11	2	3	4	3	8	4	6
1268	GCF_900041295	Neisseria meningitidis	2842STDY5880987	19400	GCA_900041295.1	2.14261	Scaffold	41	3	6	9	5	9	6	9
1269	GCF_900041305	Neisseria meningitidis	2842STDY5881213	19400	GCA_900041305.1	2.16049	Scaffold	34	8	10	5	4	5	3	8
1270	GCF_900041315	Neisseria meningitidis	2842STDY5881220	19400	GCA_900041315.1	2.11338	Scaffold	167	2	7	6	17	16	18	8
1271	GCF_900041325	Neisseria meningitidis	2842STDY5881294	19400	GCA_900041325.1	2.156	Scaffold	136	27	6	9	3	9	6	16
1272	GCF_900041335	Neisseria meningitidis	2842STDY5881350	19400	GCA_900041335.1	2.17662	Scaffold	1286	11	5	168	8	11	4	21

Table 15. Continued.

#	GeneBank Accn. Number	Organism/Name	Strain	Clade ID	Assembly	Size (Mb)	Assembly Level	PubMLST profile							
								ST	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
1273	GCF_900041345	Neisseria meningitidis	2842STDY5881364	19400	GCA_900041345.1	2.16982	Scaffold	42	10	6	9	5	9	6	9
1274	GCF_900041355	Neisseria meningitidis	2842STDY5881371	19400	GCA_900041355.1	2.12747	Scaffold	11	2	3	4	3	8	4	6
1275	GCF_900041365	Neisseria meningitidis	2842STDY5881400	19400	GCA_900041365.1	2.13108	Scaffold	11	2	3	4	3	8	4	6
1276	GCF_900041375	Neisseria meningitidis	2842STDY5881424	19400	GCA_900041375.1	2.15279	Scaffold	34	8	10	5	4	5	3	8
1277	GCF_900041385	Neisseria meningitidis	2842STDY5881439	19400	GCA_900041385.1	2.1023	Scaffold	2680	2	3	7	217	8	5	2
1278	GCF_900041395	Neisseria meningitidis	2842STDY5881452	19400	GCA_900041395.1	2.15453	Scaffold	1374	3	7	9	5	9	22	9
1279	GCF_900042395	Neisseria meningitidis	2842STDY5881620	19400	GCA_900042395.1	2.15779	Scaffold	41	3	6	9	5	9	6	9
1280	GCF_900043225	Neisseria meningitidis	2842STDY5881637	19400	GCA_900043225.1	2.12504	Scaffold	11	2	3	4	3	8	4	6
1281	GCF_900043235	Neisseria meningitidis	2842STDY5881628	19400	GCA_900043235.1	2.12261	Scaffold	153	2	3	7	2	34	5	2
1282	GCF_900043245	Neisseria meningitidis	2842STDY5881736	19400	GCA_900043245.1	2.15663	Scaffold	482	3	6	9	17	9	6	9
1283	GCF_900043255	Neisseria meningitidis	2842STDY5881060	19400	GCA_900043255.1	2.1116	Scaffold	2631	3	6	9	9	9	6	206
1284	GCF_900043265	Neisseria meningitidis	2842STDY5881098	19400	GCA_900043265.1	2.12926	Scaffold	2704	2	3	4	3	8	214	6
1285	GCF_900043275	Neisseria meningitidis	2842STDY5881340	19400	GCA_900043275.1	2.15679	Scaffold	41	3	6	9	5	9	6	9
1286	GCF_900043285	Neisseria meningitidis	2842STDY5881704	19400	GCA_900043285.1	2.16876	Scaffold	41	3	6	9	5	9	6	9
1287	GCF_900043295	Neisseria meningitidis	2842STDY5881721	19400	GCA_900043295.1	2.15256	Scaffold	41	3	6	9	5	9	6	9
1288	GCF_900043305	Neisseria meningitidis	2842STDY5881752	19400	GCA_900043305.1	2.17138	Scaffold	259	4	10	5	40	6	3	8
1289	GCF_900043315	Neisseria meningitidis	2842STDY5881397	19400	GCA_900043315.1	2.20674	Scaffold	41	3	6	9	5	9	6	9
1290	GCF_900043325	Neisseria meningitidis	2842STDY5881453	19400	GCA_900043325.1	2.11159	Scaffold	1372	2	3	7	2	8	6	2
1291	GCF_900043335	Neisseria meningitidis	2842STDY5881455	19400	GCA_900043335.1	2.16962	Scaffold	259	4	10	5	40	6	3	8
1292	GCF_900043505	Neisseria meningitidis	2842STDY5881432	19400	GCA_900043505.1	2.20817	Scaffold	3608	3	116	9	5	5	22	9
1293	GCF_900044275	Neisseria meningitidis	2842STDY5881338	19400	GCA_900044275.1	2.17765	Scaffold	269	4	10	15	9	8	11	9
1294	GCF_900044295	Neisseria meningitidis	2842STDY5881445	19400	GCA_900044295.1	2.15539	Scaffold	42	10	6	9	5	9	6	9
1295	GCF_900044705	Neisseria meningitidis	2842STDY5881562	19400	GCA_900044705.1	2.24921	Scaffold	44	9	6	9	9	9	6	9
1296	GCF_900044715	Neisseria meningitidis	2842STDY5881261	19400	GCA_900044715.1	2.15596	Scaffold	1374	3	7	9	5	9	22	9
1297	GCF_900044725	Neisseria meningitidis	2842STDY5881531	19400	GCA_900044725.1	2.13104	Scaffold	1287	2	3	4	17	8	4	6
1298	GCF_900046735	Neisseria meningitidis	2842STDY5881404	19400	GCA_900046735.1	2.16209	Scaffold	40	3	6	9	5	9	22	9
1299	GCF_900047055	Neisseria meningitidis	2842STDY5881345	19400	GCA_900047055.1	2.15099	Scaffold	34	8	10	5	4	5	3	8
1300	GCF_900047065	Neisseria meningitidis	2842STDY5881467	19400	GCA_900047065.1	2.25876	Scaffold	41	3	6	9	5	9	6	9
1301	GCF_900047985	Neisseria meningitidis	2842STDY5881584	19400	GCA_900047985.1	2.16582	Scaffold	41	3	6	9	5	9	6	9
1302	GCF_900047995	Neisseria meningitidis	2842STDY5881676	19400	GCA_900047995.1	2.16395	Scaffold	32	4	10	5	4	6	3	8
1303	GCF_900048005	Neisseria meningitidis	2842STDY5881005	19400	GCA_900048005.1	2.16236	Scaffold	4065	8	10	5	4	5	3	7
1304	GCF_900048015	Neisseria meningitidis	2842STDY5881205	19400	GCA_900048015.1	2.1873	Scaffold	22	11	5	18	8	11	24	21
1305	GCF_900048025	Neisseria meningitidis	2842STDY5881228	19400	GCA_900048025.1	2.13216	Scaffold	1466	6	5	173	13	5	24	17
1306	GCF_900048435	Neisseria meningitidis	2842STDY5881339	19400	GCA_900048435.1	2.24002	Scaffold	5445	3	6	9	5	9	22	2

Table 16. List of genomes used for the gene detection tests.

#	GenBank Accn. Number	Organism	Description	NCBI Taxa Id
1	CP010829.1	*Shigella sonnei*	Shigella sonnei strain FORC_011, complete genome	624
2	CP024466.1	*Shigella dysenteriae*	Shigella dysenteriae strain BU53M1 chromosome, complete genome	622
3	CP027027.1	*Shigella dysenteriae*	Shigella dysenteriae strain E670/74 chromosome, complete genome	622
4	HE616528.1	*Shigella sonnei*	Shigella sonnei 53G main chromosome, complete genome	216599
5	NC_000915.1	*Helicobacter pylori*	Helicobacter pylori 26695 chromosome, complete genome	85962
6	NC_000921.1	*Helicobacter pylori*	Helicobacter pylori J99, complete genome	85963
7	NC_000962.3	*Mycobacterium tuberculosis*	Mycobacterium tuberculosis H37Rv, complete genome	83332
8	NC_002163.1	*Campylobacter jejuni*	Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819 chromosome, complete genome	192222
9	NC_002516.2	*Pseudomonas aeruginosa*	Pseudomonas aeruginosa PAO1 chromosome, complete genome	208964
10	NC_002695.1	*Escherichia coli*	Escherichia coli O157:H7 str. Sakai, complete genome	386585
11	NC_002745.2	*Staphylococcus aureus*	Staphylococcus aureus subsp. aureus N315 DNA, complete genome	158879
12	NC_002755.2	*Mycobacterium tuberculosis*	Mycobacterium tuberculosis CDC1551, complete genome	83331
13	NC_002758.2	*Staphylococcus aureus*	Staphylococcus aureus subsp. aureus Mu50 DNA, complete genome	158878
14	NC_002946.2	*Neisseria gonorrhoeae*	Neisseria gonorrhoeae FA 1090 chromosome, complete genome	242231
15	NC_002951.2	*Staphylococcus aureus*	Staphylococcus aureus subsp. aureus COL, complete genome	93062
16	NC_003028.3	*Streptococcus pneumoniae*	Streptococcus pneumoniae TIGR4, complete genome	170187
17	NC_003098.1	*Streptococcus pneumoniae*	Streptococcus pneumoniae R6 chromosome, complete genome	171101
18	NC_003197.2	*Salmonella enterica*	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2, complete genome	99287
19	NC_003912.7	*Campylobacter jejuni*	Campylobacter jejuni RM1221, complete genome	195099
20	NC_004337.2	*Shigella flexneri*	Shigella flexneri 2a str. 301 chromosome, complete genome	198214
21	NC_004631.1	*Salmonella enterica*	Salmonella enterica subsp. enterica serovar Typhi Ty2, complete genome	209261
22	NC_004741.1	*Shigella flexneri*	Shigella flexneri 2a str. 2457T, complete genome	198215
23	NC_006511.1	*Salmonella enterica*	Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150, complete genome	295319
24	NC_007146.2	*Haemophilus influenzae*	Haemophilus influenzae 86-028NP, complete genome	281310
25	NC_007384.1	*Shigella sonnei*	Shigella sonnei Ss046, complete genome	300269
26	NC_007606.1	*Shigella dysenteriae*	Shigella dysenteriae Sd197 chromosome, complete genome	300267
27	NC_008258.1	*Shigella flexneri*	Shigella flexneri 5 str. 8401, complete genome	373384
28	NC_008463.1	*Pseudomonas aeruginosa*	Pseudomonas aeruginosa UCBPP-PA14, complete genome	208963
29	NC_008533.1	*Streptococcus pneumoniae*	Streptococcus pneumoniae D39, complete genome	373153
30	NC_008787.1	*Campylobacter jejuni*	Campylobacter jejuni subsp. jejuni 81-176, complete genome	354242
31	NC_009525.1	*Mycobacterium tuberculosis*	Mycobacterium tuberculosis H37Ra, complete genome	419947

Table 16. Continued.

#	GenBank Accn. Number	Organism	Description	NCBI Taxa Id
32	NC_009648.1	*Klebsiella pneumoniae*	Klebsiella pneumoniae subsp. pneumoniae MGH 78578, complete genome	272620
33	NC_009656.1	*Pseudomonas aeruginosa*	Pseudomonas aeruginosa PA7, complete genome	381754
34	NC_010410.1	*Acinetobacter baumannii*	Acinetobacter baumannii str. AYE, complete genome	509173
35	NC_010554.1	*Proteus mirabilis*	Proteus mirabilis strain HI4320, complete genome	529507
36	NC_010611.1	*Acinetobacter baumannii*	Acinetobacter baumannii ACICU, complete genome	405416
37	NC_011035.1	*Neisseria gonorrhoeae*	Neisseria gonorrhoeae NCCP11945, complete genome	521006
38	NC_011283.1	*Klebsiella pneumoniae*	Klebsiella pneumoniae 342, complete genome	507522
39	NC_011586.2	*Acinetobacter baumannii*	Acinetobacter baumannii AB0057, complete genome	480119
40	NC_011750.1	*Escherichia coli*	Escherichia coli IAI39 chromosome, complete genome	585057
41	NC_012731.1	*Klebsiella pneumoniae*	Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044 DNA, complete genome	484021
42	NC_014121.1	*Enterobacter cloacae*	Enterobacter cloacae subsp. cloacae ATCC 13047 chromosome, complete genome	716541
43	NC_015663.1	*Klebsiella aerogenes*	Enterobacter aerogenes KCTC 2190 chromosome, complete genome	1028307
44	NC_016514.1	*Enterobacter cloacae*	Enterobacter cloacae EcWSU1, complete genome	1045856
45	NC_016809.1	*Haemophilus influenzae*	Haemophilus influenzae 10810 genome	862964
46	NC_017022.1	*Enterococcus faecium*	Enterococcus faecium Aus0004, complete genome	1155766
47	NC_017382.1	*Helicobacter pylori*	Helicobacter pylori 51, complete genome	290847
48	NC_017451.1	*Haemophilus influenzae*	Haemophilus influenzae R2866, complete genome	262728
49	NC_017634.1	*Escherichia coli*	Escherichia coli O83:H1 str. NRG 857C chromosome, complete genome	685038
50	NC_017731.1	*Providencia stuartii*	Providencia stuartii MRSN 2154, complete genome	1157951
51	NC_017960.1	*Enterococcus faecium*	Enterococcus faecium DO chromosome, complete genome	333849
52	NC_018079.1	*Enterobacter cloacae*	Enterobacter cloacae subsp. dissolvens SDM, complete genome	1104326
53	NC_020181.1	*Klebsiella aerogenes*	Enterobacter aerogenes EA1509E complete genome	935296
54	NC_020207.1	*Enterococcus faecium*	Enterococcus faecium NRRL B-2354, complete genome	1104325
55	NC_020211.1	*Serratia marcescens*	Serratia marcescens WW4, complete genome	435998
56	NC_020418.1	*Morganella morganii*	Morganella morganii subsp. morganii KT, complete genome	1124991
57	NC_022000.1	*Proteus mirabilis*	Proteus mirabilis BB2000, complete genome	1266738
58	NC_022240.1	*Neisseria gonorrhoeae*	Neisseria gonorrhoeae MS11, complete genome	528354
59	NC_022347.1	*Campylobacter coli*	Campylobacter coli CVM N29710, complete genome	1273173
60	NC_022660.1	*Campylobacter coli*	Campylobacter coli 15-537360, complete genome	1358410
61	NZ_CP008920.1	*Providencia stuartii*	Providencia stuartii strain ATCC 33672, complete genome	588
62	NZ_CP011574.1	*Klebsiella aerogenes*	Klebsiella aerogenes strain CAV1320 chromosome, complete genome	548
63	NZ_CP017054.1	*Providencia stuartii*	Providencia stuartii strain BE2467 chromosome, complete genome	588

Table 16. Continued.

#	GenBank Accn. Number	Organism	Description	NCBI Taxa Id
64	NZ_CP017671.1	*Providencia rettgeri*	Providencia rettgeri strain RB151, complete genome	587
65	NZ_CP019977.1	*Campylobacter coli*	Campylobacter coli strain aerotolerant OR12, complete genome	195
66	NZ_CP023505.1	*Morganella morganii*	Morganella morganii strain FDAARGOS_365 chromosome, complete genome	582
67	NZ_CP023965.1	*Proteus vulgaris*	Proteus vulgaris strain FDAARGOS_366 chromosome, complete genome	585
68	NZ_CP026046.1	*Morganella morganii*	Morganella morganii strain FDAARGOS_63 chromosome, complete genome	582
69	NZ_CP026050.1	*Serratia marcescens*	Serratia marcescens strain FDAARGOS_65 chromosome, complete genome	615
70	NZ_CP026062.1	*Proteus mirabilis*	Proteus mirabilis strain FDAARGOS_81 chromosome, complete genome	584
71	NZ_HG326223.1	*Serratia marcescens*	Serratia marcescens subsp. marcescens Db11, complete genome	273526

APPENDIX B. SUPPLEMENTARY DATA FOR CHAPTER 3

B.1 WebSTing data dictionary

Table 17. Table SEC_ROLE. Table to hold all role information for users of the system.

Column Name	Type	Description
id	Long	Primary Index Key
authority	String	Type of role for the system (eq: ROLE_ADMIN, ROLE_USER)

Table 18. Table SEC_ USER. Table to hold all user information for the system.

Column Name	Type	Description
id	Long	Primary Index Key
username	String	unique name for users
password	String	password for the system (Encrypted MD5)
enabled	Boolean	flag for enabling and disabling a user
accountExpired	Boolean	flag for expired account
accountLock	Boolean	flag to set locking status
password_expired	Boolean	Filed used to flag for an expired password

Table 19. Table SEC_ USER_ROLE. Table to hold the relationship between users and roles.

Column Name	Type	Description
id	Long	Primary Index Key
sec_user_id	Long	user id
sec_role_id	Long	role id

Table 20. Table ORGAN_TYPE_SCHEME. Table to hold the relationship between users and roles.

Column Name	Type	Description
id	Long	Primary Index Key
sting_id	Long	Unique Id used to retrieve records from the STing system
display_name	String	Organism name used for display to users
scheme	String	Scheme type for organisms
orig_name	String	Original name of organisms
date_create	Date	The date the record was created
last_updated	Date	The date this record was last updated

Table 21. Table UPLOADED_FILES. Table to hold queue for files uploaded to the webserver.

Column Name	Type	Description
id	int	Primary Index Key
name	String	Actual file name
size	bigInt	Size of the file
save_time	Datetime	time in which the file is saved and recorded in DB
file_1_uploaded	Boolean	Is the 1 file uploaded
file_1_zip	Boolean	Is the file a zip file
file_2_uploaded	Boolean	Is the 2 files uploaded
file_2_zip	Boolean	Is the 2 file a zip file
nmb	Int	Schema type number to run in STing
organism_name	String	The actual organism name

Table 22. Table PROCESSING_FILES. Table to hold information about the record as it is being processed through STing.

Column Name	Type	Description
`id`	int	Primary Index Key
`Name`	String	Actual accession name
`start_time`	Datetime	Actual start time the STing process started to process this accession
`end_time`	Datetime	Actual end time the STing process ended in processing this accession
`line_type`	String	Type of information displayed
`status`	String	Status during the STing process
`st`	Int	Sequence Type (ST)
`total_kmers`	Int	Total *k*-mers processed
`total_reads`	Int	Total reads processed
`organism_name`	String	The actual organism name
`nmb`	Int	Schema type number to run in STing

Table 23. Table SAMPLE_ALLELES. A table of generated alleles produced from STing for an accession.

Column Name	Type	Description
`id`	int	Primary Index Key
`processing_file_id`	Int	Foreign key from Processing_files table
`locus`	String	Locus name
`allele`	Int	Allele number
`positions`	Int	Allele length
`norm_count`	String	Normalized *k*-mer count
`coverage`	Int	Allele coverage
`mean_kmer_depth`	Int	Average *k*-mer depth
`sd_kmer_depth`	String	*k*-mer depth standard deviation
`image_file`	String	Absolute file path to PNG image generated from STing for this allele

Table 24. Table RETRIEVE_ACCESSION_FILES. Table to hold the queue for accession numbers that need to be searched.

Column Name	Type	Description
id	int	Primary Index Key
accesion_name	String	Actual accession name
nmb	Int	Schema type number to run in STing
organism_name	String	The actual organism name
status	String	Status of the file retrieval process

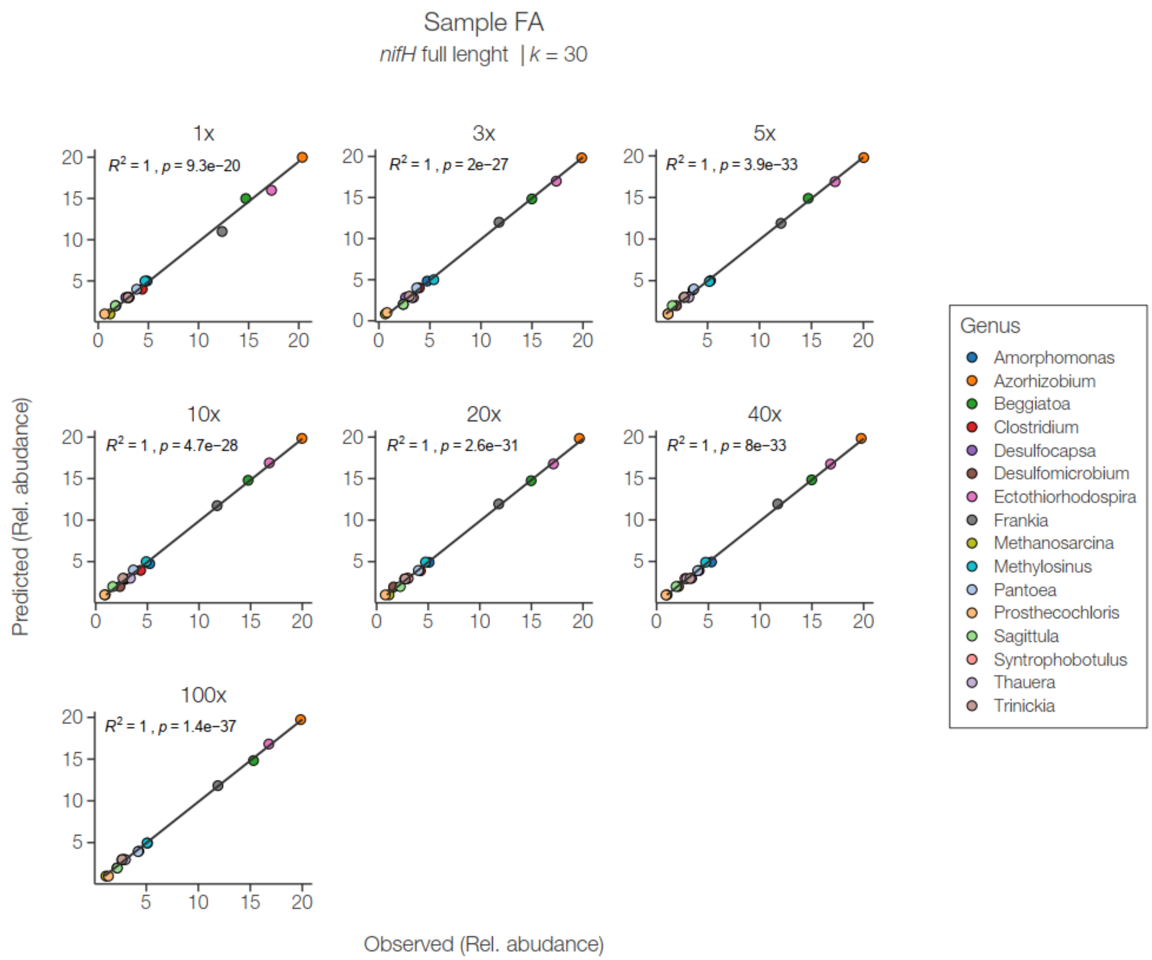


Figure 34. Comparison of the predicted and observed relative abundance of the sample FA. Plots show the relative abundance at the genus level calculated after classifying the simulated read sets with STing (predicted) as a function of the actual relative abundance of the corresponding simulated dataset (observed) for the simulated sample FA at different sequencing depths.

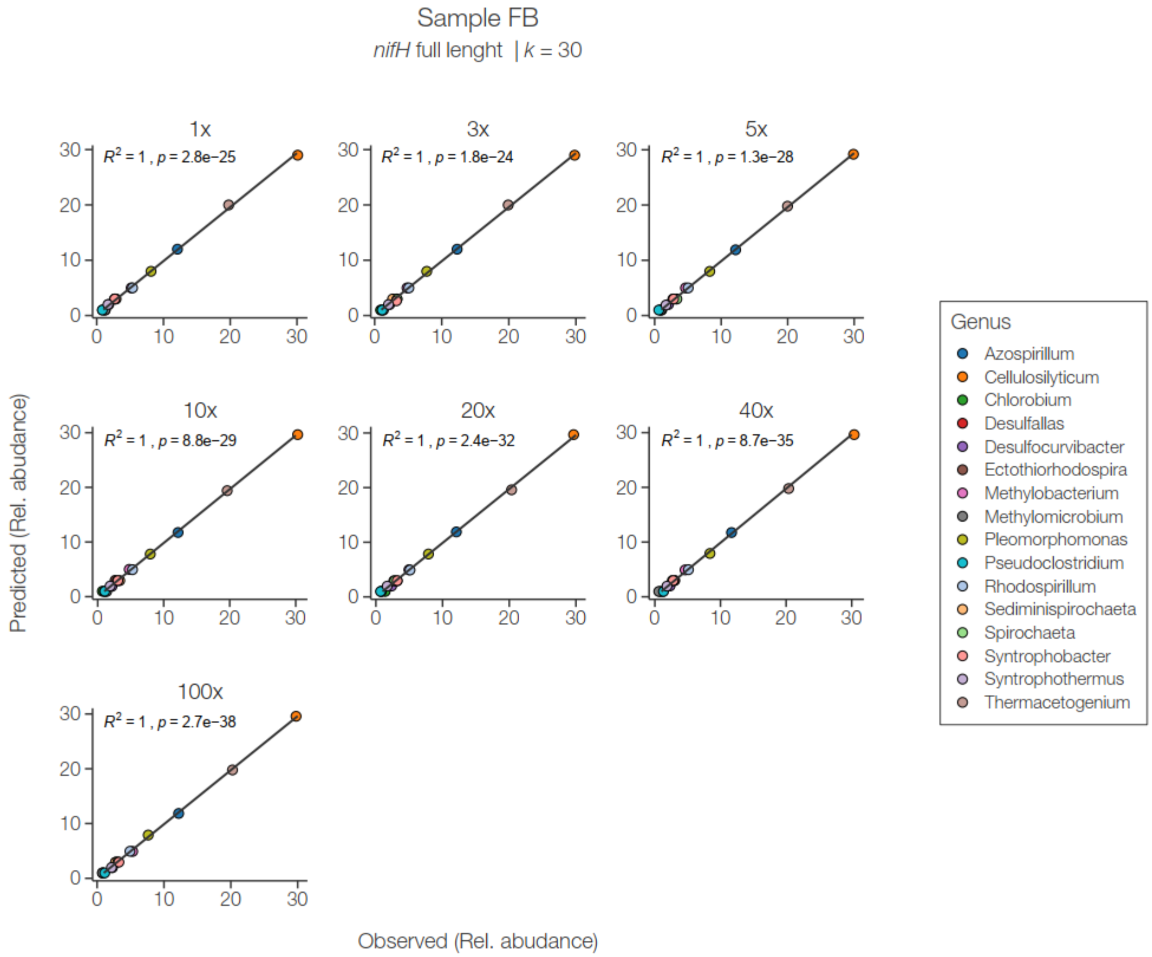


Figure 35. Comparison of the predicted and observed relative abundance of the sample FB. Plots show the relative abundance at the genus level calculated after classifying the simulated read sets with STing (predicted) as a function of the actual relative abundance of the corresponding simulated dataset (observed) for the simulated sample FB at different sequencing depths.

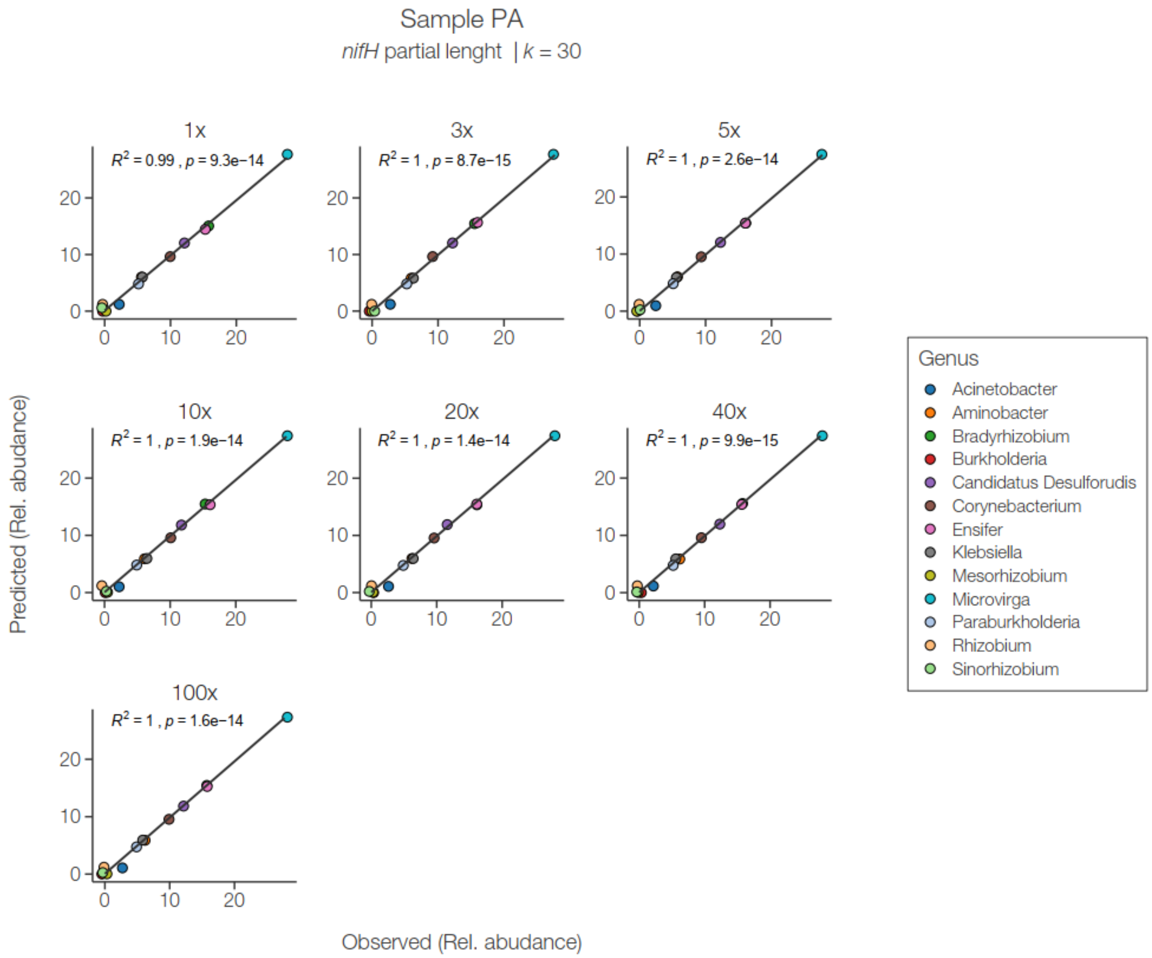


Figure 36. Comparison of the predicted and observed relative abundance of the sample PA. Plots show the relative abundance at the genus level calculated after classifying the simulated read sets with STing (predicted) as a function of the actual relative abundance of the corresponding simulated dataset (observed) for the simulated sample PA at different sequencing depths.

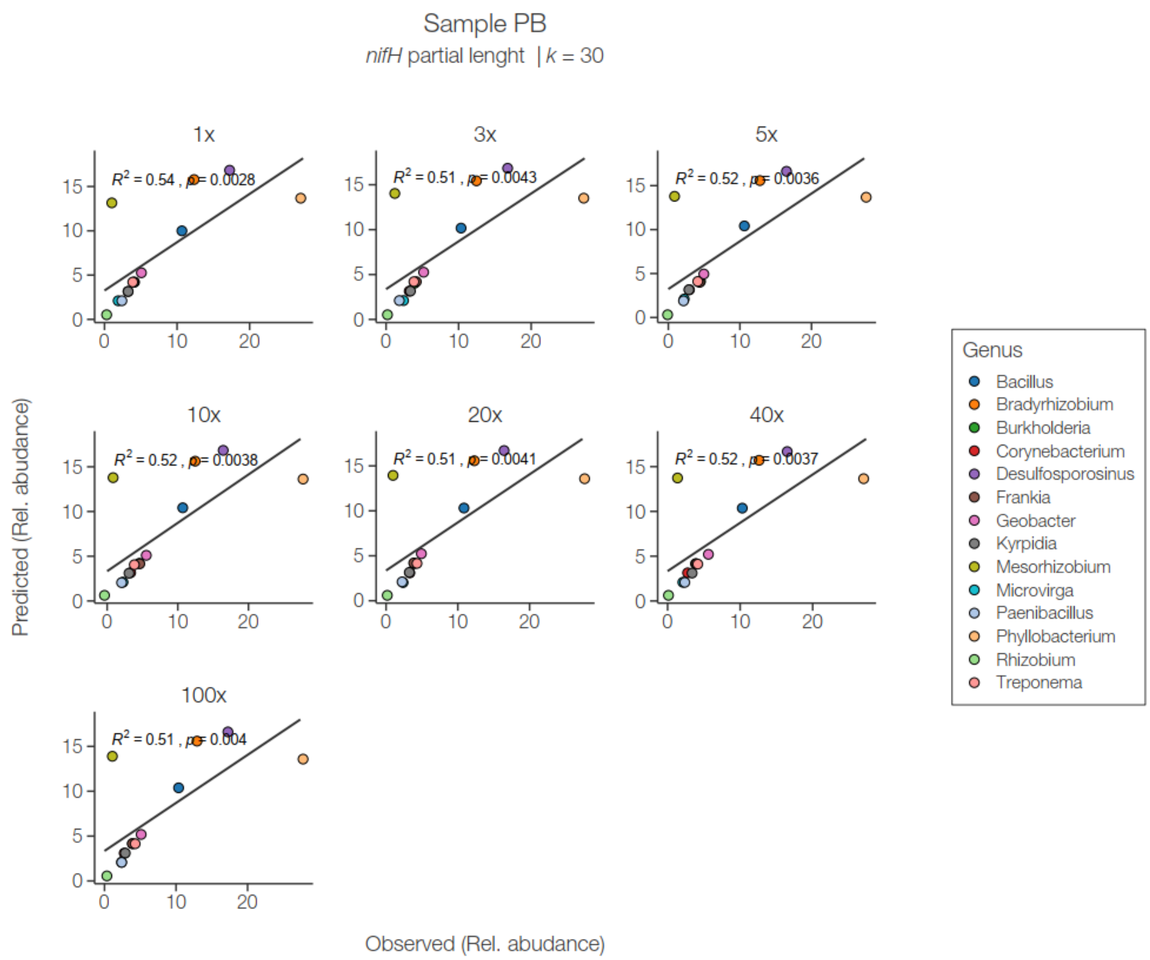


Figure 37. Comparison of the predicted and observed relative abundance of the sample PB. Plots show the relative abundance at the genus level calculated after classifying the simulated read sets with STing (predicted) as a function of the actual relative abundance of the corresponding simulated dataset (observed) for the simulated sample PB at different sequencing depths.

REFERENCES

Abouelhoda, M.I., Kurtz, S. and Ohlebusch, E. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms* 2004;2(1):53--86.

Afgan, E., *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;46(W1):W537-W544.

Angiuoli, S.V., *et al.* CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 2011;12:356.

Audano, P. and Vannberg, F. KAnalyze: a fast versatile pipelined k-mer toolkit. *Bioinformatics* 2014;30(14):2070-2072.

Audano, P.A., Ravishankar, S. and Vannberg, F.O. Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. *Bioinformatics* 2018;34(10):1659-1665.

Aziz, R.K., *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;9:75.

Bankevich, A., *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19(5):455-477.

Camacho, C., *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.

Camacho, C., *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10(1):421.

Connor, T.R., *et al.* CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb Genom* 2016;2(9):e000086.

Espitia-Navarro, H.F., Chande, A. T., Nagar S. D., Smith, H., Jordan, I. K., Rishishwar, L. STing: accurate and ultrafast genomic profiling with exact sequence matches. *BioRXiv* 2019.

Espitia-Navarro, H.F., Rishishwar, L., Mayer, L. W., Jordan, I. K. Bioinformatics. In: Budowle, B.a.S., S. and Morse, S., editor, *Microbial Forensics*. Academic Press; 2019.

Espitia, H., *et al.* A method of sequence typing with in silico aptamers from a next generation sequencing platform. In.; 2017.

Feijao, P., *et al.* MentaLiST - A fast MLST caller for large MLST schemes. *Microbial genomics* 2018;4(2):1--8.

Feijao, P., *et al.* MentaLiST - A fast MLST caller for large MLST schemes. *Microb Genom* 2018.

Ferres, I. and Iraola, G. MLSTar: automatic multilocus sequence typing of bacterial genomes in R. *PeerJ* 2018;6:e5098.

Fricke, W.F. and Rasko, D.A. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat Rev Genet* 2014;15(1):49-55.

Gaby, J.C., *et al.* Diazotroph Community Characterization via a High-Throughput nifH Amplicon Sequencing and Analysis Pipeline. *Appl Environ Microbiol* 2018;84(4).

Gould, L.H., *et al.* Recommendations for diagnosis of shiga toxin--producing Escherichia coli infections by clinical laboratories. *MMWR Recomm Rep* 2009;58(RR-12):1-14.

Gupta, A., Jordan, I.K. and Rishishwar, L. stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics* 2017;33(1):119-121.

Gupta, S.K., *et al.* ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother* 2014;58(1):212-220.

Huang, W., *et al.* ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;28(4):593-594.

Hunt, M., Mather, A.E. and Sanchez-Buso, L. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial Genomics* 2017;3(10).

Hunt, M., *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017;3(10):e000131.

Inouye, M., *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine* 2014;6(11):90.

Inouye, M., *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014;6(11):90.

Jia, B., *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2017;45(D1):D566-D573.

Joensen, K.G., *et al.* Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *J Clin Microbiol* 2014;52(5):1501-1510.

Jolley, K.A., *et al.* Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 2012;158(Pt 4):1005-1015.

Jolley, K.A. and Maiden, M.C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;11:595.

Katz, L.S., *et al.* Meningococcus genome informatics platform: a system for analyzing multilocus sequence typing data. *Nucleic Acids Res* 2009;37(Web Server issue):W606-611.

Katz, L.S., *et al.* Evolutionary dynamics of Vibrio cholerae O1 following a single-source introduction to Haiti. *MBio* 2013;4(4).

Krampis, K., *et al.* Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinformatics* 2012;13:42.

Lakin, S.M., *et al.* MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res* 2017;45(D1):D574-D580.

Larsen, M.V., *et al.* Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol* 2012;50(4):1355-1361.

Liu, B., *et al.* VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* 2019;47(D1):D687-D692.

Maiden, M.C., *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 1998;95(6):3140-3145.

Maiden, M.C., *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95(6):3140--3145.

Maiden, M.C., *et al.* MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* 2013;11(10):728-736.

Maiden, M.C.J., *et al.* MLST revisited: The gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology* 2013;11(10):728--736.

McDonald, D., *et al.* The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* 2012;1(1):7.

Page, A.J., *et al.* Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microb Genom* 2017;3(8):e000124.

Page, A.J. and Keane, J.A. Rapid multi-locus sequence typing direct from uncorrected long reads using Krocus. *PeerJ* 2018;6:e5233.

Page, A.J., Taylor, B. and Keane, J.A. Multilocus sequence typing by blast from de novo assemblies against PubMLST. *J Open Source Softw* 2016;1:118-111.

Parsons, B.D., *et al.* Detection, Characterization, and Typing of Shiga Toxin-Producing Escherichia coli. *Front Microbiol* 2016;7:478.

Pritchard, L. 2014. run_MLST.py. https://github.com/widdowquinn/scripts/blob/master/bioinformatics/run_MLST.py

Reinert, K., *et al.* The SeqAn C++ template library for efficient sequence analysis: A resource for programmers. *J Biotechnol* 2017;261:157-168.

Silva, M., *et al.* chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb Genom* 2018.

Spellerberg, I.F. and Fedor, P.J. A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index. *Global Ecol Biogeogr* 2003;12(3):177-179.

Stucki, D., *et al.* Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *J Infect Dis* 2015;211(8):1306-1316.

Tacconelli, E., *et al.* Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis* 2018;18(3):318-327.

Tewolde, R., *et al.* MOST: a modified MLST typing tool based on short read sequencing. *PeerJ* 2016;4:e2308.

Thomsen, M.C., *et al.* A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance. *PLoS One* 2016;11(6):e0157718.

Wang, X., Jordan, I.K. and Mayer, L.W. A phylogenetic perspective on molecular epidemiology. In: Tang, Y.-W., *et al.*, editors, *Molecular Medical Microbiology (Second Edition)*. Chennai, India: Elsevier; 2015. p. 517-536.

Weber, N., *et al.* Nephele: a cloud platform for simplified, standardized and reproducible microbiome data analysis. *Bioinformatics* 2018;34(8):1411-1413.